XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT- 8 – INFORMAÇÃO E TECNOLOGIA

GERAÇÃO AUTOMÁTICA DE METADADOS: ESTUDO DE CASO UTILIZANDO A TÉCNICA DE INDEXAÇÃO AUTOMÁTICA ESTATÍSTICA COM A FERRAMENTA ANNIF

AUTOMATIC METADATA GENERATION: CASE STUDY USING THE AUTOMATIC STATISTICAL INDEXING TECHNIQUE WITH THE ANNIF TOOL

Jean Carlos Borges Brito. UNB.

Dalton Martins. UNB.

Modalidade: Trabalho Completo

Resumo: Esta pesquisa apresenta um estudo de caso com a ferramenta ANNIF, executando a geração automática de metadados através da técnica de indexação automática estatística e aprendizagem de máquina, utiliza algoritmo baseado em regras para extrair valores de metadados dos recursos de informação. O objetivo do trabalho é elaborar um *framework* para utilização da ferramenta. Criou-se um corpus de conhecimento com 52 artigos da Base Brasileira de Ciência da Informação (BRAPCI), utilizando como vocabulário controlado o Tesauro Brasileiro em Ciência da Informação (TBCI). Após o processo de treinamento do modelo realizou-se teste preliminar de indexação automática estatística sobre uma Tese Completa armazenada no Repositório Institucional da Universidade de Brasília (RiUnB) gerando a recomendação de assuntos/descritores. Os termos atribuídos pelo ANNIF foram comparados com as palavras-chave da tese da RiUnB, obtendo boa similaridade. Conclui-se que o uso do ANNIF, utilizando a técnica de indexação automática estatística contribuiu para automatização da tarefa, obtendo desempenho satisfatório.

Palavras-Chave: Geração Automática de Metadados. Aprendizagem de Máquina. Indexação. ANNIF.

Abstract: This research presents a case study with the ANNIF tool, performing the automatic generation of metadata through the technique of automatic statistical indexing and machine learning, using a rule-based algorithm to extract metadata values from information resources. The objective of the work is to develop a framework for using the tool. A corpus of knowledge was created with 52 articles from the Brazilian Information Science Base (BRAPCI), using the Brazilian Thesaurus in Information Science (TBCI) as a controlled vocabulary. After the model training process, a preliminary test of automatic statistical indexing was carried out on a Complete Thesis stored in the Institutional Repository of the University of Brasília (RiUnB), generating the recommendation of subjects/descriptors. The terms assigned by the ANNIF were compared with the keywords of the RiUnB thesis, obtaining good similarity. It is concluded that the use of ANNIF, using the technique of automatic statistical indexing, contributed to the automation of the task, achieving satisfactory performance.

Keywords: Automatic Generation of Metadata. Machine Learning. Indexing. ANNIF.

1 INTRODUÇÃO

A busca e a recuperação de objetos de pesquisa nos repositórios digitais tem sido impactada devido a diversos problemas relacionados aos metadados e a representação da informação nesses acervos.

Mooers (1951) afirma que um usuário potencial da informação é capaz de converter sua necessidade de informações em uma lista de referências para documentos armazenados e que contém informações úteis. Neste contexto, os metadados desempenham um papel essencial, pois descrevem os dados, facilitam sua compreensão e corroboram na eficácia da catalogação e na sua obtenção.

De acordo com Pomerantiz (2015), metadado indica algo que está além dos dados, sendo uma declaração sobre esses dados. Haynes (2018) corrobora com o entendimento de que o metadado tem a função de facilitar o entendimento dos relacionamentos e evidenciar a utilidade das informações obtida dos dados.

Os gestores de repositório digitais são responsáveis por uma quantidade enorme de metadados relacionados a diferentes tipos de documentos que geralmente são indexados por títulos, assuntos e descritores para que possam ser recuperados posteriormente (SUOMINEN, 2019). Entretanto, nem todos os usuários de sistemas de biblioteca e repositórios digitais executam a entrada correta e completa de metadados, o que dificulta a recuperação do objeto de pesquisa.

Polfreman et *al.* (2008) discorrem que sem os metadados apropriados, os recursos permanecem ocultos e sem utilização, causando desperdício de investimento. O autor também enfatiza que a baixa qualidade ou metadados inexistentes são igualmente eficazes para tornar os recursos inutilizáveis, pois sem ele um recurso é essencialmente invisível dentro de um repositório ou arquivo morto e, portanto, permanece desconhecido e inacessível.

Crystal e Land (2003) discorrem que para criar metadados para um milhão de documentos deveriam ser alocados 60 empregados/ano para realizar essa tarefa. O processo manual de geração de metadados de documentos é um trabalho árduo, oneroso e dependendo do volume de dados é humanamente impossível de ser realizado. Além disso, considerando que o conceito de documento e suas possibilidades de expressão midiática se expandem de forma significativa na era da web, torna-se proibitivo imaginar que a

catalogação dos documentos seguirá continuamente sendo realizada apenas de forma manual.

1.1 REPOSITÓRIOS DIGITAIS: EM FOCO A BUSCA E RECUPERAÇÃO DA INFORMAÇÃO

Gusmão et *al.* (2017) discorre que no paradigma sócio-tecnológico contemporâneo, os indivíduos produzem informações e as ofertam em diversos ambientes de informação digital, tendo como objetivo a promoção da interação entre esses indivíduos, dos indivíduos com as instituições e entre as instituições.

De acordo com Pavão et *al.* (2015), as tecnologias de informação e comunicação (TIC) são cada dia mais utilizadas pelas instituições de ensino e pesquisa com o intuito de oferecer informações sobre sua coleção de documentos, preservando seu conteúdo informacional em meio digital, fazendo uso dos repositórios institucionais.

Shintaku e Vidotti (2016) discorrem que a disponibilização das informações através da internet (web 2.0) nas últimas décadas, propiciou o advento da mudança do físico para o digital, surgindo assim, os ambientes de informação digital, que nesta tipologia se encontram os repositórios digitais (GUSMÃO et al., 2017).

Araújo, Maia e Vechiato (2018) discorrem que o termo "repositório" se origina da palavra em latim repositorium e significa "um local onde os objetos poderiam ser armazenados e coletados". Os autores enfatizam que os repositórios digitais surgiram como uma resposta espontânea às dificuldades e custos associados para divulgação dos periódicos científicos, pela evolução da TIC e da necessidade de armazenar e disseminar o patrimônio intelectual de várias instituições.

O Tesauro Brasileiro da Ciência da Informação (TBCI), através de uma Nota Explicativa (NE), conceitua o termo "repositórios digitais" como:

> Mecanismos para administrar, armazenar e preservar conteúdos informacionais em formato eletrônico, e que podem ter como foco um assunto (repositórios temáticos) ou a produção científica de uma instituição (repositórios institucionais). Muitos permitem o acesso universal e gratuito a seus conteúdos, que variam de acordo com a política de cada instituição. São coleções digitais de documentos de interesse para a pesquisa científica e, no caso dos institucionais, representam a sua memória científica (PINHEIRO e FERREZ, 2014, p. 195).

Nesse contexto, podemos representar os repositórios digitais de acordo com a seguinte estrutura esquemática:

Figura 1 – Estrutura esquemática – repositórios digitais

Fonte: Elaborado pelo autor.

Sanchez e Vechiato (2017) também fazem a distinção entre repositórios digitais temáticos e institucionais. Enquanto o primeiro cobre determinada área temática de conhecimento com escopo bem definido, o segundo é composto por sistemas de informação que armazenam, preservam, divulgam e fornecem acesso a produção técnica e intelectual das instituições e comunidades científicas.

Pavão et *al.* (2015) enfatizam a importância da padronização, normalização e enriquecimento dos metadados para fortalecimento da qualidade dos registros nos repositórios digitais. Estes são atributos extremamente importantes que garantem a descrição e a identificação do documento, auxiliando na obtenção dos resultados em buscas executadas nos sistemas automatizados com o objetivo de aumentar a satisfação do usuário. Para isso, os metadados são uma ferramenta essencial para a melhoria da representação da informação.

Cerrao e Castro (2018) afirmam que para utilizar os repositórios digitais (sejam eles temáticos ou institucionais) de forma adequada e funcional, deve-se realizar ações de representação e descrição dos recursos informacionais de forma padronizada para auxiliar na busca e recuperação da informação (RI. A RI da informação nos repositórios institucionais:

> [...] parte de um princípio de que a informação foi registrada e armazenada de forma adequada, seguindo padrões de catalogação e uso de metadados e com conteúdo e estrutura de informação muito bem delimitada e separada, baseada em conceitos que se preocupam com a recuperação da informação, como o uso de estrutura e formatos de representação da informação previamente estudados (SANTAREM SEGUNDO, 2020, p. 163).

Compreende-se que os repositórios digitais são estruturas de organização da informação e possui relação direta com a busca e a recuperação da informação. Ressalta-se, portanto, a importância de investigações que enfatizem a representação da descrição dos

recursos informacionais de forma automática, suportando de forma adequada a recuperação da informação e a satisfação dos usuários desses repositórios digitais.

Polfreman et *al.* (2008) elenca seis técnicas para a geração de metadados: colheita de metatags, extração de conteúdo, mineração de textos e dados, *folksonomia* ou marcação social, geração automática de metadados extrínsecos e indexação ou classificação automática. Esta pesquisa vai aprofundar nesta última técnica.

1.2 INDEXAÇÃO OU CLASSIFICAÇÃO AUTOMÁTICA

De acordo com Bandim e Correa (2019), a indexação é "um dos processos de análise documentária realizada com a finalidade de determinar para cada documento um conjunto de palavras-chave ou assuntos". Esses autores afirmam que a organização e a recuperação da informação são materializadas via processo de indexação e sua automatização tem sido adotada, visando a aplicação em um volume cada vez mais crescente de artigos científicos e da necessidade de elaboração de índices de busca que facilitem sua recuperação.

Essa técnica envolve o uso de aprendizado de máquina e algoritmos baseados em regras para extrair valores de metadados dos próprios recursos de informação, conforme Park e Brenza (2015). No entanto, estes autores afirmam que a técnica também abrange o mapeamento de termos de metadados extraídos para vocabulários controlados. Enfatizam que os pesquisadores utilizam algoritmos de classificação e agrupamento para extrair metadados relevantes dos textos. Demonstram também que são utilizadas estatísticas de frequência de termo em oposição à sua relativa não frequência em documentos relacionados.

De acordo com Gil Leiva (2009) o alcance da qualidade na recuperação da informação é obtido através do processo de indexação de documentos.

Suominem (2019) afirma que a indexação manual é uma tarefa intelectual que demanda bastante tempo, sendo que muitos destes artefatos estão em formato digital, tornando possível a automatização do trabalho de indexação a partir do texto completo ou certas partes de documentos, como títulos e resumos.

A indexação automática pode se dividir em dois tipos: por atribuição ou por extração, conforme Silva e Correia (2020). A extração de termos dos textos dos documentos, fornecendo-lhe pesos e selecionando aqueles mais expressivos, representando seu conteúdo temático denomina-se de indexação automática por extração (LANCASTER, 2004, p. 18-19). Enquanto que atribuir termos ao documento através de outra fonte, tal como o emprego de

termos extraídos de um vocabulário controlado, denomina-se de indexação automática por atribuição. Lancaster (2004) discorre que um vocabulário controlado é uma lista de termos autorizados com uma forma de estrutura semântica (significado), que controlam sinônimos, distinguem homógrafos e agrupam termos afins, podendo ser divididos em três tipos: esquemas de classificação bibliográfica, lista de cabeçalhos de assuntos e tesauros. Pinheiro e Ferrez (2014, p.9) conceitua tesauros como:

> [...] instrumentos de organização do conhecimento, ou melhor, como **linguagens documentárias utilizadas no processo de indexação**, são listas estruturadas de termos e suas relações, onde cada um deve representar um único conceito ou ideia, de forma a **orientar indexadores** e usuários, levando-os de uma ideia ao termo que melhor a expresse. Desta forma, tesauros de diversos campos do saber vêm sendo publicados para facilitar a recuperação da informação. [grifo nosso]

As linguagens documentárias são conjuntos de termos descritos no texto e seus vínculos, utilizadas no processo de indexação e denota sua representação mental, sua imagem. Ao executar os algoritmos que utilizam técnicas de aprendizagem de máquina, as linguagens documentárias orientam os indexadores, fornecendo a compreensão dos termos dentro da lista controlada e estruturada, auxiliando na análise de assuntos e na recuperação de documentos e publicações nas mais variadas áreas de conhecimento humano.

O uso de vocabulário controlado através de tesauros melhora o processo de indexação automática de documentos, pois auxilia as ferramentas computacionais no fornecimento de termos comumente utilizados.

1.3 ANNIF – FERRAMENTA DE INDEXAÇÃO AUTOMÁTICA ESTATÍSTICA

De acordo com Lappalainen et *al.* (2021) a ferramenta ANNIF tem despertado interesse em muitas organizações e a experiência das primeiras implementações na Biblioteca Nacional da Finlândia tem sido promissoras. O ponto de partida é a seleção adequada do vocabulário de assuntos adequado e um corpus de conhecimento para ensinar os modelos de aprendizado de máquina e a combinação de algoritmos para diferentes abordagens para obtenção dos melhores resultados.

ANNIF[1] é uma solução mantida pela Biblioteca Nacional da Finlândia de código aberto e baseada em microserviço. A ferramenta foi apresentada pelo Sr. Osma Suominen, especialista em sistemas de informação na Biblioteca Nacional da Finlândia na *47th LIBER Annual Conference* em 2018 na França, Sessão 10, com o título "*Annif: Feeding your subject indexing robot with bibliographic metadata*".

De acordo com Suominen (2018), o protótipo inicial foi desenvolvido em 2017 e vem sofrendo atualizações constantes e que estão sendo disponibilizadas no repositório do GitHub, sob a licença Apache 2.0. Esta solução foi concebida para executar a indexação automática de assuntos e classificação a partir de diferentes coleções de documentos, dentre artigos científicos, dissertações, livros antigos digitalizados, e-books e arquivos (SUOMINEN, 2018). ANNIF é multilíngue e suporta qualquer vocabulário de assunto tanto em formato *Simple Knowledge Organization System* (SKOS) ou *Tab Separated Values* (TSV). É possível acessar sua interface através de linha de comando, formato web ou através de microserviço REST-API. Essa solução combina o uso de ferramentas de processamento de linguagem natural e aprendizagem de máquina, incluindo os algoritmos Maui, Omikuji, fastText e Gensim.

Suominen (2019, p. 21-22) destaca que "as funcionalidades de indexação são gerenciadas por diferentes algoritmos que podem ser usados separadamente ou combinados nos chamados conjuntos". Dessa forma, cada algoritmo pode ser implementado como módulos separados e novos algoritmos podem ser adicionados posteriormente. A instalação do ANNIF pode conter um ou vários projetos independentes, onde cada um deles especifica uma série de parametrizações, tais como o *backend*, o idioma e o vocabulário de indexação. Cada projeto é limitado a um único idioma, mas a indexação multilíngue poderá ser executada definindo vários projetos, sendo cada um deles por idioma. Em cada projeto é definido um número, geralmente grande de assuntos, que espelham a ideia ou significado de um vocabulário de indexação. Os assuntos são criados a partir de um corpus que é extraído dos registros de metadados presentes e/ou documentos indexados. No ANNIF poderá haver vários *backends* independentes que auxiliam com sugestões de assuntos. Estes *backends* podem ser integrados ao ANNIF, ou serviços externos que poderão ser consultados via *Application Program Interface* (API), sendo que um projeto pode executar vários *backends* e combinar os seus resultados de análise.

[1] Zenodo DOI: https://doi.org/10.5281/zenodo.2578948

ANNIF utiliza tipos diferentes de corpus de documentos e de assuntos, normalmente é utilizado um tesauro, uma classificação ou lista de cabeçalhos de assuntos. A ferramenta não se preocupa com a estrutura interna do vocabulário de assuntos, precisando apenas compreender as URIs e os rótulos de preferência (termos ou descritores) de cada um dos assuntos ou classes ou conceitos. A ferramenta também pode utilizar um ou vários corpus de documentos para treinar modelos baseados em estatística ou aprendizagem de máquina e também avalia o desempenho desses modelos. Uma curiosidade interessante é que possui suporte a dois formatos de corpus, sendo um mais adequado para aqueles documentos robustos ou compridos (textos completos ou resumos extensos) e outro adequado para textos pequenos, tais como o título ou palavras-chave de um artigo.

2 DESENVOLVIMENTO

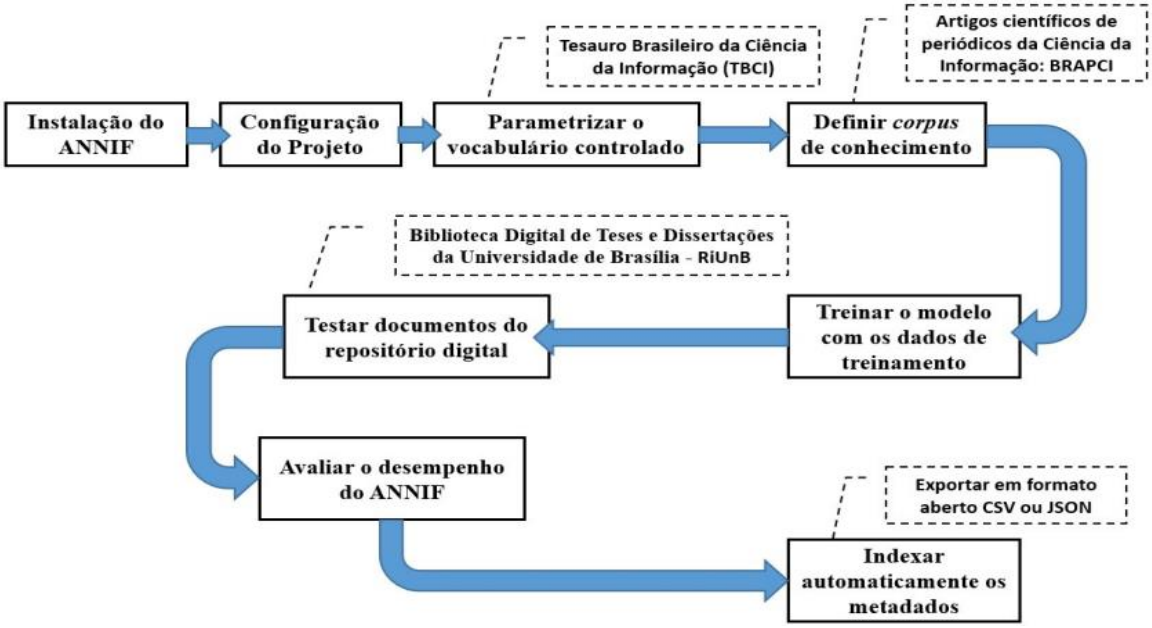
A pesquisa procura atingir um determinado objetivo e, para que isso se concretize, é executada uma investigação científica que está sujeita a uma série de processos, atividades e procedimentos cognitivos e técnicos que denominamos de métodos científicos. A investigação neste artigo será realizada através de estudo de caso, sendo caracterizada pela pesquisa em profundidade e exaustiva da geração automática de metadados, fornecendo vasto conhecimento em detalhes dos fenômenos relativos à (RI), mais precisamente à indexação automática. De acordo com Yin (2005, p. 32), o estudo de caso "é uma investigação empírica que investiga um fenômeno contemporâneo dentro do seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos".

A RI é um fenômeno que é estudado há décadas, mas o tema se torna contemporâneo quando se observa que nos últimos 20 anos, o volume e variedade da informação aumentou consideravelmente, sendo necessário mecanismos automatizados para auxiliar os gestores de repositórios digitais na organização desses acervos. A pesquisa será pragmática, no sentido de estudar e aplicar a ferramenta tecnológica denominada ANNIF para geração automática de metadados, baseado em dados de um repositório digital real visando melhorar a busca e recuperação desse repositório. O uso desta ferramenta está fundamentado nas características apontadas por Suominen, Inkinen e Lehtinen (2022): arquitetura modular, ferramenta multilíngue, independente de vocabulário de indexação e classificação, suporte a diferentes algoritmos de forma flexível e adaptável a diferentes situações, integrável a outros sistemas

através de suas interfaces (linha de comando – CLI, Web e Rest API) e seu fornecimento e disponibilização é realizada através de código aberto.

A seguir, é descrito o *framework* proposto nesta pesquisa, descrevendo os passos para execução do ANNIF que serão aplicados no estudo de caso:

Figura 2 – *Framework*: Passos de execução do ANNIF



Fonte: Elaborado pelo autor.

2.1 INSTALAÇÃO DO ANNIF

A máquina virtual (VM) com a imagem do tutorial ANNIF, possui o Sistema Operacional Ubuntu (64-bit), Memória RAM de 6GB e 2 (dois) processadores. Por ser ambiente Linux, a configuração do ANNIF e a execução dos testes são facilitados para quem tem mais familiaridade com sistemas Unix/Linux.

O *download* da *VirtualBox* pode ser realizado através do link <https://annif.org/download/>.

2.2 CONFIGURAÇÃO DO PROJETO

O ANNIF requer a configuração de um ou mais projetos para sua utilização. O projeto consiste em um conjunto de definições tais como: sua identificação (ID), descrição, linguagem/idioma, *backend*/algoritmo, vocabulário controlado, analisador. Nesta etapa realizou-se as parametrizações de configuração no arquivo “projects.cfg”, podendo ser visualizadas no exemplo a seguir:

Figura 3 – Configurando o projeto no ANNIF


```
[brapci]
name=Projeto Tesauro Brasileiro da Ciencia da Informacao
language=pt
backend=mllm
vocab=vc-tbci
analyzer=snowball(portuguese)
```

Fonte: Elaborado pelo autor.

Nesta investigação utilizou-se o algoritmo *Maui-like Lexical Matching* (MLLM) para realização da indexação automática, uma vez que esse *backend* não necessita de um corpus de volume robusto. Este algoritmo, segundo Suominen (2021), é uma reimplementação em Python de vários conceitos utilizados no Maui, com algumas adaptações. Esse algoritmo necessita ser treinado com um número relativamente pequeno (centenas ou milhares) de documentos indexados manualmente para que o algoritmo escolha a combinação correta de heurísticas que supra os melhores resultados em uma determinada coleção de documentos. Após a execução da configuração, basta rodar o comando "annif list-projects" para mostrar a lista de entrada definida para o projeto.

2.3 PARAMETRIZAR O VOCABULÁRIO CONTROLADO

Terceiro passo foi parametrizar o vocabulário controlado de assuntos selecionado. Neste estudo de caso será o Tesauro Brasileiro da Ciência da Informação (TBCI *online*, mantido pela Universidade Estadual de Londrina (UEL). O arquivo deve estar no formato: <uri>*tab*"Termo da TBCI", conforme Figura a seguir:

Figura 4 – Formato do arquivo do vocabulário controlado para carga no ANNIF

<http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1323&/marc> MARC
<http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=590&/buscas-de-informacao> Buscas de Informação
<http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=24&/33-servicos-de-informacao> Serviços de Informação
<http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1374&/governo-eletronico> Governo Eletrônico
<http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=28&/41-inteligencia-competitiva> Inteligência Competitiva
<http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=188&/economia-da-informacao> Economia da Informação
<http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=923&/informacao-estrategica> Informação Estratégica

Fonte: Elaborador pelo autor.

Após a criação do arquivo do vocabulário controlado (vc-tbci.tsv), realizou-se a carga no projeto denominado "brapci", através do comando:

annif loadvoc brapci data-sets/brapci/vc-tbci.tsv

2.4 DEFINIR UM *CORPUS* DE CONHECIMENTO

Quarto passo foi treinar o modelo utilizando os dados de treinamento, observando o idioma atribuído ao projeto e neste caso foi o português. O *corpus* de assunto e documentos foi construído sobre os dados de título, resumo e palavras-chave de 52 artigos extraídos da Base de Dados em Ciência da Informação (BRAPCI). O formato para este arquivo foi: "Título do Artigo". "Resumo do Artigo". "Palavras-chave".tab<uri>.

Depois do *corpus* de assunto criado (corpus-brapci.tsv), executou-se o treinamento do modelo através do comando:

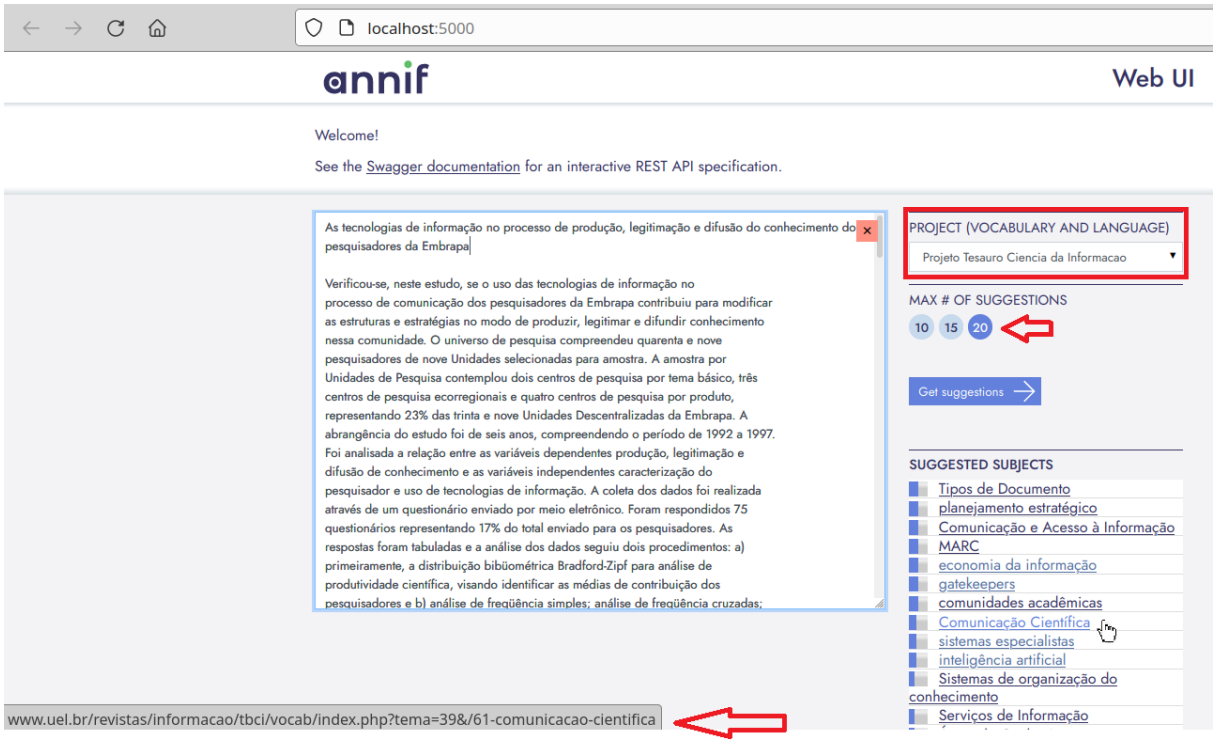
annif train brapci data-sets/brapci/corpus-brapci.tsv

2.5 TESTAR DOCUMENTO DO REPOSITÓRIO DIGITAL - RiUnB

Quinto passo será testar os documentos do repositório digital RiUnB, solicitando sugestões de assuntos ao ANNIF utilizando o *backend*/algoritmo parametrizado ao projeto.

Acessou-se a RiUnB em 18/04/2022, "FCI – Faculdade de Ciência da Informação" e executada pesquisa pelo assunto "Tecnologia da Informação e Comunicação". Selecionou-se o documento de Tese com o Título "As tecnologias de informação no processo de produção, legitimação e difusão do conhecimento dos pesquisadores da Embrapa" de Souza (1999). Utilizou-se o texto da tese completa para testar a indexação automática com 20 sugestões de assuntos pelo ANNIF. O teste foi executado via API REST contendo como vocabulário controlado, o Tesauro da TBCI parametrizado ao projeto:

Figura 5 – Sugestão de Assuntos via API REST com ANNIF, uso de Projeto com Tesauro TBCI



Fonte: Adaptado de Suominen (2019).

Após o carregamento do texto completo, seleção do projeto e a definição do número máximo de sugestões, clicou-se no botão "*Get suggestion*". Automaticamente, o ANNIF realizou a sugestão de assuntos, ranqueando os resultados e disponibilizando *link* direto ao termo do Tesauro da TBCI, neste exemplo "Comunicação Científica", conforme Figura 6. Dessa forma, o usuário poderá verificar os termos genéricos, específicos e relacionados com a sugestão de assunto fornecido pelo ANNIF.

Figura 6 – Redirecionamento para o TBCI ao clicar no assunto sugerido pelo ANNIF

Fonte: Tesauro Brasileiro de Ciência da Informação – TBCI (UEL, 2014).

O ANNIF realizou a sugestão de 20 assuntos classificando em ordem de significância, sendo neste caso: 1) Tipo de Documento; 2) Planejamento Estratégico; 3) Comunicação e acesso à informação; 4) MARC; 5) Economia da Informação; 6) *Gatekeepers*; 7) Comunidades Acadêmicas; 8) Comunicação Científica; 9) Sistemas Especialistas; 10) Inteligência Artificial; 11) Sistemas de Organização do Conhecimento; 12) Serviços de Informação; 13) Áreas de Conhecimento; 14) Inovação; 15) Buscas de Informação; 16) Tecnologias da Informação e Comunicação; 17) Informação Estratégica; 18) Representação da Informação; 19) Organização do Conhecimento; e 20) Sociedade da Informação.

2.6 AVALIAR O DESEMPENHO DO ANNIF

É possível avaliar o desempenho do ANNIF e obter uma série de medidas estatísticas tais como: *Precision* (doc avg), *Recall* (doc avg), F1 *Score* (doc avg), *Precision* (conc avg), *Recall* (conc avg), F1 *Score* (conc avg), *Precision* (microavg), *Recall* (microavg), F1 *Score* (microavg), NDCG, NDCG@5, NDCG@10, Precision@1, Precision@3, Precision@5, *LRAP, True positives, False positives, False negatives e Documents evaluated*. Entretanto, neste estudo o enfoque

foi na proposição de *framework* definindo uma sequência de passos para geração automática de metadados através da técnica de indexação automática estatística e aprendizagem de máquina, utilizando a ferramenta ANNIF.

Está em curso o aprofundamento de pesquisas para ampliação do corpus de conhecimento, assim como a aplicação de outros *backends*/algoritmos executados isoladamente e em conjunto, extraindo o potencial de cada um para geração automática de metadados. Os resultados, assim como a avaliação de desempenho do ANNIF, serão objeto de publicações futuras.

3 CONSIDERAÇÕES FINAIS

A solução ANNIF se demonstrou nos testes preliminares uma solução robusta e que pode auxiliar no processo de indexação automática de metadados, através do uso de um conjunto de algoritmos distintos trabalhando em conjunto, com intuito de auxiliar os gestores de acervos digitais a organizar melhor a sua coleção de documentos. Utilizou-se neste estudo apenas o algoritmo/*backend* MLLM, mas o ANNIF pode utilizar:

a) algoritmos léxicos que combinam termos de um documento para termos contidos em um vocabulário controlado, executando comparação com uso de poucos dados de treinamento (Ex.: Maui, YAKE, MLLM, STWFSA); e

b) algoritmos associativos que são aqueles que aprendem quais conceitos estão correlacionados e com quais termos nos documentos, baseado nos dados de treinamento. A abordagem associativa precisa de muito mais dados de treinamento para cobrir cada assunto (Ex.: TF-IDF, *fastText* e *Vowpal Wabbit*).

Maiores investigações devem ser realizadas executando a indexação automática de assuntos na RiUnB com o uso conjunto de algoritmos léxicos e associativos.

O *framework* proposto se mostrou adequado para a realização das atividades planejadas, descrevendo: os passos necessários para criação do vocabulário controlado, utilizando os termos da TBCI; construção do corpus de conhecimento com os 52 (cinquenta e dois) artigos da BRAPCI; Treinamento do Modelos e sugestão de assuntos.

REFERÊNCIAS

ARAÚJO, Aline Karoline da Silva; MAIA, Flávio Henrique; VECHIATO, Fernando Luiz. **Encontrabilidade da informação em repositórios digitais: um estudo de caso na Biblioteca Digital de Monografias da UFNR**. Rev. Inf. na Soc. Contemp., Natal, RN, v.2, n1, jan./jun., 2018.

BANDIM, Marcio Aercio Silva; CORREA, Renato Fernandes. **Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação**. Transinformação, v.31, e180004, 2019. http://dx.doi.org/10.1590/2318-0889201931e180004

CERRAO, Natália Gallo; CASTRO, Fabiano Ferreira de. **Repositórios institucionais das Universidades Federais brasileiras: análise da representação da informação**. Informação & Tecnologia (ITEC), Marília/João Pessoa, v.5, n.1, jan./jun. 2018.

CRYSTAL, Abe; LAND, Paula. ***Metadata and Search: Global Corporate Circle DCMI 2003 Workshop***. 2003. Disponível em <http://www.dublincore.org/groups/corporate/Seattle/> (acessado em 05 de abril de 2022).

GIL LEIVA, Isidoro. ***Manual de indización: teoría y práctica***. Gijón: Trea, 2009

GUSMÃO, Felipe Carvalho Marinho; SILVA, Mayane Paulino de Brito e; PEREIRA, Giuliane Monteiro; LIMA, Izabel França de; OLIVEIRA, Henry Poncio Cruz de. **Elementos de arquitetura da informação no Repositório Eletrônico Institucional da UFPB**. Rev. Inf. na Soc. Contemp., Natal, RN, Número Especial, 2017.

HAYNES, David. ***Metadata for Information Management and Retrieval: understanding metadata and its use.*** 2ª Edição, London: Facet Publishing, 2018.

LANCASTER, Frederick Wilfrid. **Indexação e resumos: teoria e prática**. 2. ed. Brasília: Briquet de Lemos Livros. 452p., 2004.

LAPPALAINEN, Mikko; HULKKONEN, Juha; INKINEN, Juho; KALLIO, Aleksi; LEHTINEN, Mona; KOSKELA, Markus; SJÖBERG, Mats; SUOMINEN, Osma; YETUKURI, Laxmana. ***Automaattisen sisällönkuvailun ohjelmiston rakentaminen – case Annif***. Signum, vol. 53, nº 4, 14–20, 2021.

MOOERS, Calvin Northrup. ***Zatocoding applied to mechanical organization of knowledge. American Documentation***, v.2, n.1, 1951, p.20-32.

PARK, Jung-ran; BRENZA, Andrew. ***Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art***. Information Technology and Libraries, Volume 34, Ed. 3, p. 22-42, Chicago, USA, 2015.

PAVÃO, Caterina Groposo; COSTA, Janise Borges da; FERREIRA, Manuela Klanovicz; HOROWITZ, Zaida. **Metadados e repositórios institucionais: uma relação indissociável para a qualidade da recuperação e visibilidade da informação**. PontodeAcesso, Salvador, v.9, n.2, p.103-116, dez. 2015.

PINHEIRO, Lena Vania; FERREZ, Helena Dodd. **Tesauro Brasileiro de Ciência da Informação**. Rio de Janeiro; Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), 2014.

POLFREMAN, Malcolm; BROUGHTON, Vanda; WILSON, Andrew. ***Metadata Generation for Resource Discovery. JISC***, 2008. Disponível em http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/autometgen.aspx

POMERANTZ, Jeffrey. ***Metadata***. Cambridge, MA: MIT Press, 2015.

SANCHEZ, Fernanda Alves; VECHIATO, Fernando Luiz. **Encontrabilidade da Informação em repositórios digitais: um enfoque nos repositórios institucionais da USP, UNESP e UNICAMP**. XVIII Enancib 2017, 23 a 27 de outubro de 2017, Marília, São Paulo, 2017.

SANTARÉM SEGUNDO, José Eduardo. **Representação Iterativa: um modelo para repositórios digitais**. 2010. 224 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010. Disponível em: <http://www.marilia.unesp.br/Home/PosGraduacao/CienciadaInformacao/Dissertacoes/santaremsegundo_je_do_mar.pdf>. Acesso em: 28 fev. 2022.

SHINTAKU, Milton; VIDOTTI, Silvana Aparecida Borsetti Gregorio. **Bibliotecas e repositórios no processo de publicação digital**. Biblos: Revista do Instituto de Ciências Humanas e da Informação, v. 30, n.1, 2016.

SILVA, Sâmela Rouse de BRITO; CORREA, Renato Fernandes. **Sistemas de Indexação Automática por atribuição: uma análise comparativa**. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 25, p. 01-15, 2020. Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2020.e70740

SUOMINEN, Osma***. Annif: Feeding your subject indexing robot with bibliographic metadata***. Liber's 47th Annual Conference in Lille, France, Data Enhancements in the Service of Research Libraries, session 10, 2018.

SUOMINEN, Osma***. Annif: DIY Automated Subject Indexing Using Multiple Algorithms***. Liber Quarterly, vol. 29, 2019.

SUOMINEN, Osma; INKINEN, Juho; LEHTINEN, Mona. ***Annif and Finto AI: Developing and Implementing Automated Subject Indexing***. JLIS.it, vol. 13, nº 1, january, 2022.

UEL. **Tesauro Brasileiro de Ciência da Informação**. TBCI adaptado de Lena Vânia Ribeiro Pinheiro. Universidade Estadual de Londrina. 2014. Disponível em < http://www.uel.br/revistas/informacao/tbci/vocab/>. Acesso em 20 ago. 2022.

YIN, Robert. K. **Estudo de caso: planejamento e métodos**. Porto Alegre, RS: Bookman, 2005