Recuperação Automatizada de Documentos Científicos por Meio da Análise de Planos de Ensino

Flavio Izo, UFES/IFES, fizo@ifes.edu.br, https://orcid.org/0000-0001-7189-0387
Elias Oliveira, UFES, elias@lcad.inf.ufes.br, https://orcid.org/0000-0003-2066-7980

Resumo: O plano de ensino é um instrumento pedagógico que descreve a ementa, os objetivos, os conteúdos programáticos, o processo avaliativo e a bibliografia básica dos componentes curriculares presentes na grade escolar. Por isso, o plano de ensino serve como guia para que os docentes e discentes busquem fontes de informações para complementar o processo de aprendizagem. Neste trabalho, apresentamos um sistema que emprega técnicas de Processamento de Linguagem Natural (PLN) para recuperar documentos relevantes aos conteúdos delineados nos planos de ensino, oferecendo, assim, recursos de apoio aos estudos. Nossos experimentos envolveram 15 planos de ensino, nos quais recuperamos materiais complementares de bancos de artigos científicos e patentes para 46.43% dos conteúdos identificados.

Palavras-chave: Recuperação da informação, Processamento de Linguagem Natural, Plano de Ensino, Reconhecimento de Entidades Nomeadas, Inteligência Artificial.

Automated Retrieval of Scientific Documents through Teaching Plan Analysis

Abstract: The teaching plan is a pedagogical instrument that describes the syllabus, objectives, programmatic contents, the evaluation process, and the primary bibliography of the curricular components present in the school curriculum. Therefore, the teaching plan serves as a guide for teachers and students to seek sources of information to complement the learning process. In this work, we present a system that uses Natural Language Processing (NLP) techniques to retrieve documents relevant to the content outlined in the teaching plans, thus offering resources to support studies. Our experiments included 15 teaching plans, and we retrieved complementary materials from databases of papers and patents for 46.43% of the identified content.

Keywords: Information Retrieval, Natural Language Processing, Teaching Plan, Named Entity Recognition, Artificial Intelligence.

1. Introdução

As Instituições de Ensino (IE) utilizam diversos documentos para se estruturar e se organizar de maneira clara e objetiva, alinhado com os padrãos educacionais regulamentados pelo Ministério da Educação e Cultura (MEC). Projeto Pedagógico, Calendário Escolar, Plano de Ensino, Planos de Aula, Boletim Escolar e Estatuto são alguns dos documentos institucionais que servem para garantir o funcionamento, a organização e a prestação dos serviços educacionais.

Os planos de ensino contemplam os conteúdos essenciais ao processo de ensino e aprendizagem dos alunos, concentrando-se na definição dos objetivos do componente curricular e contribuindo para a avaliação do progresso discente. Assim, o plano de ensino representa a maneira pela qual o professor organiza e sistematiza as atividades didáticas, a fim de facilitar o processo de aprendizagem (FERREIRA; REHFELDT; SILVA, 2020; ASTUTIK; ROSID, 2018; AHMAD; YAACOB, 2018) e também servir de referência para o fornecimento de materiais de apoio para recuperação (ASTUTIK; ROSID, 2018).

Para uma aprendizagem eficaz, é fundamental consultar fontes além das referenciadas no plano de ensino. A publicação de materiais complementares é uma

estratégia que pode contribuir para a melhoria efetiva da qualidade de ensino (ALFERES; MAINARDES, 2014).

De acordo com Mooers (1951 apud SARACEVIC, 1992, p. 44), a Recuperação da Informação (RI) "engloba os aspectos intelectuais de descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação". No entanto, a RI pode utilizar conceitos e técnicas de Inteligência Artificial (IA) para aperfeiçoar a recuperação de maneira mais inteligente para o usuário (CERI *et al.*, 2013).

Hoje em dia, a disponibilidade de informações é abundante, especialmente com a ascensão da internet. Com o avanço das tecnologias da informação, houve também o aumento de informações disponíveis para consulta. A Internet tornou a busca por estas informações mais extensas, seja através de notícias, mídias sociais, bibliotecas online, artigos etc. Recuperar informações personalizadas é uma tarefa crucial para a seleção de dados em situações em que há uma grande quantidade de informações disponíveis (WU *et al.*, 2021). No entanto, esta não é uma tarefa trivial. Investigar dados e recuperar informações manualmente é dispensioso e desafiador. Portanto, é necessário empregar um método automatizado para processar os documentos de forma eficiente (CAN; KABADAY1, 2021).

Com a crescente expansão e popularização da IA, surgem novos desafios que a área de Ciência da Informação deve abordar. É essencial realizar pesquisas e implementar práticas que permitam avanços em diversos campos de estudo, com destaque para a representação e a recuperação da informação (CONEGLIAN, 2018).

O objetivo deste trabalho é verificar se a análise automatizada dos planos de ensino na área de química pode identificar informações pertinentes para a pesquisa em documentos de patentes e artigos científicos, contribuindo para a recuperação eficaz de materiais científicos e enriquecendo o processo de estudo dos alunos.

O componente curricular química foi selecionado pois é uma área que está presente em diversos tipos de documentos, sejam artigos científicos, patentes, bulas de remédios, descrições de produtos, informações sobre tratamento químico entre outros. Dentre os documentos de consulta, os artigos científicos e as patentes foram selecionados pois são exemplos de fontes de aprendizagem que fornecem suporte para a produção de conhecimento (IZO *et al.*, 2023).

A primeira contribuição deste trabalho reside na apresentacão de uma metodologia para a recuperação de documentos que sirvam como materiais complementares aos componentes curriculares escolares. A segunda é que, considerando documentos de patentes, pode-se aproximar os planos de ensino de novidades científicas e tecnológicas registradas. Além disso, abre-se um caminho para medir a distância entre os planos de ensino e aquilo que as instituições de ensino estejam dependentes. Importante destacar que, embora tenha sido utilizado artigos científicos e patentes como exemplos de materiais complementares, o sistema é capaz de recuperar outros tipos de documentos.

Este artigo está organizado da seguinte forma. Na Seção 2, alguns trabalhos relacionados são brevemente revisados. A metodologia utilizada para o desenvolvimento da abordagem proposta está na Seção 3. Na Seção 4 são descritos os experimentos, os resultados encontrados e as limitações do trabalho. As considerações finais e os trabalhos futuros estão na Seção 5.

2. Trabalhos Relacionados

Nesta Seção, destacam-se alguns estudos que investigaram o uso do plano de ensino como um recurso de apoio ao processo educacional.

O trabalho de Astutik e Rosid (2018) descreve uma aplicação de sistema que simplifica a gestão do plano de ensino para os professores, ao mesmo tempo em que auxilia a IE na avaliação da eficácia da aprendizagem e na adequação do diário de ensino. Os autores automatizaram o processo de criação do plano de ensino utilizando o método *Framework for the Application of System Thinking* (FAST) e o modelo responsivo *Twitter Bootstrap*. O processo de validação usou testes de caixa preta para averiguar a estabilidade do sistema e contaram com professores, com o responsável pela área pedagógica e com alunos. Os resultados mostraram que as funcionalidades facilitam a gestão do plano de ensino. Observou-se que o sistema apresenta uma funcionalidade muito interessante, que possibilita a recuperação digital do plano de ensino. Essa funcionalidade elimina a necessidade de extrair texto de arquivos digitais, evitando risco de perda de dados (SOPER; FUJIMOTO; YU, 2021; KARTHIKEYAN *et al.*, 2021; HILL; HENGCHEN, 2019; PANDEY *et al.*, 2022).

O trabalho de Ahmad e Yaacob (2018) destaca a integração automática entre as atividades semanais da estrutura do plano de ensino e as atividades de *e-learning* através do *plugin Embedding Mechanism* da plataforma de aprendizagem do *Moodle*. Como resultado, os autores destacam que os educadores podem preparar o plano de ensino uma única vez e, em seguida, incorporá-lo ao ambiente de *e-learning*, eliminando a necessidade de duplicar esforços e reduzindo o tempo despendido. Esse método oferece total suporte à integração das atividades do plano de ensino na estrutura de *e-learning*, cumprindo o conceito de reutilização, sem a necessidade de preparar atividades separadas para o plano de ensino e para o *e-learning*.

O trabalho conduzido por Aburajab e Salman (2019) apresenta o desenvolvimento e a implementação de um quadro negro interativo em tempo real, projetado para integração em um sistema de tutoria baseado na *web*. A principal finalidade dessa ferramenta é facilitar a transição da experiência de ensino presencial para o ambiente *online*, permitindo que professores e alunos comuniquem-se em tempo real, tanto verbalmente quanto por meio de escrita e desenhos, em um quadro negro virtual compartilhado. O sistema foi concebido como um módulo a ser integrado em um ambiente de sistema de tutoria. Para a sua implementação, os autores utilizaram tecnologias como *Canvas HTML5* e *WebSockets*. Entre os módulos do sistema, destaca-se o *Course Manager*, que possibilita a criação dinâmica de planos de ensino voltados para cursos específicos, refletindo uma sequência de tópicos de estudo que os alunos devem seguir.

O trabalho de Zhang e Hou (2022) apresenta a geração automática de gráficos do conhecimento chinês, com base no plano de ensino educacional, para melhorar a eficiência do ensino. O sistema extrai automaticamente os pontos de conhecimento e as devidas relações, e em seguida aloca o peso do relacionamento de acordo com as regras de relacionamento e gera o gráfico de conhecimento com base no plano de ensino. Os autores utilizaram o modelo *hidden Markov* para segmentar as palavras e posteriormente utilizaram *encoder* para extrair as relações entre os conceitos chaves.

Diferentemente dos estudos citados anteriormente, a nossa proposta consiste em um método para recuperar automaticamente documentos científicos a partir de informações contidas no plano de ensino. Diversas tecnologias foram envolvidas com este objetivo, que será abordado na Seção 3. Assim, segundo nosso entendimento, nenhum dos estudos possui uma abordagem com a proposta parecida com a nossa. No

entanto, é importante salientar funcionalidades interessantes observadas nos trabalhos anteriores, como a de preenchimento do plano de ensino digital (ASTUTIK; ROSID, 2018; ABURAJAB; SALMAN, 2019), a integração entre plano de ensino e as atividades de aula (AHMAD; YAACOB, 2018) e a geração de gráficos de conhecimento (ZHANG; HOU, 2022). Nosso método, em conjunto com essas funcionalidades, e aliadas à praticidade dos planos de ensino em formato digital, possibilitam a vinculação automática de materiais científicos complementares aos conteúdos estudados pelos alunos, facilitando o acesso e a disponibilização desses recursos de maneira eficiente.

3. Metodologia

Em termos metodológicos este trabalho é classificado como uma pesquisa aplicada. Para avaliar a abordagem proposta neste trabalho, algumas etapas foram definidas. A Figura 1 mostra o fluxo das ações realizadas e que serviram para a posterior execução dos experimentos.

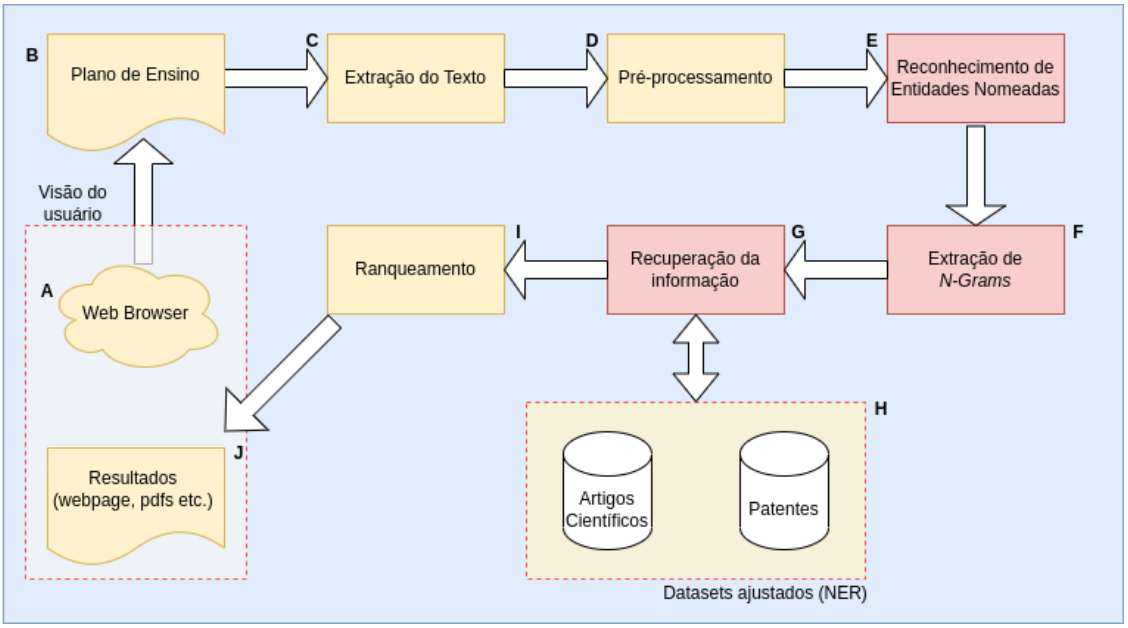

Figura 1. Fluxograma do sistema desenvolvido.

Etapas A e B: Na visão do usuário, há a disponibilidade de uma plataforma *web* onde é possível fazer o *upload* do plano de ensino e este ficará disponível para processamento. Esta funcionalidade foi desenvolvida com tecnologia de *HyperText Markup Language* – HTML5 (*Linguagem de Marcação de Hipertexto*, em tradução livre) juntamente com a linguagem *Python*.

Etapas C e D: Extração do texto do plano de ensino em formato *Portable Document Format* – PDF) através do *toolkit Apache Tika*[1]juntamente com a linguagem de programação *Python* e posterior pré-processamento do texto para eliminar *stopwords*. O *Python package Natural Language Toolkit* – NLTK[2]foi utilizado para remover as *stopwords* e linhas em branco, e também foram removidos alguns caracteres especiais.

Etapa E: Reconhecimento das Entidades Nomeadas (NER) através da *Local Grammar – LG (GROSS, 1997)* (Gramática Local, em tradução livre). As LGs são regras escritas manualmente, por meio de grafos, para reconhecer termos em um texto. Para

[1] ⟨https://tika.apache.org/⟩
[2] ⟨https://www.nltk.org⟩

esta finalidade foi utilizada a LG denominada LG*q* Plus (IZO *et al.*, 2023). A LG*q* Plus foi criada para abordar a carência de um conjunto de treinamento contendo entidades químicas previamente reconhecidas e reconhece as seguintes categorias de entidades químicas: *Classes Químicas, Compostos Químicos, Elementos Químicos, Equipamentos Químicos e Métodos Químicos*.

Etapa F: Extração de *N-Grams* para segmentar palavras existentes no plano de ensino. Foi utilizada a linguagem *Python* e algumas bibliotecas de PLN para esta atividade. Através do contexto linguístico, foi utilizado o método (*N-gram* para selecionar uma sequência *n* de palavras para serem pesquisadas nos *datasets* de artigos e patentes. Um *unigram* (*1-gram*) seleciona as palavras individualmente; o *bigram* (*2-gram*) seleciona duas sequências de palavras; o *trigram* (*3-gram*) seleciona três sequencias de palavras; e assim sucessivamente. Na sentença *"Refratômetro é um equipamento químico"*, *"Refratômetro"* é um *unigram*, *"Refratômetro é"* é um *bigram* e *"Refratômetro é um"* é um *trigram* e assim por diante. É importante ressaltar que conforme aumenta-se o valor de *n*, aumenta-se a qualidade do modelo, no entanto, o modelo cresce exponencialmente juntamente com o tamanho do vocabulário (FERREIRA; LOPES, 2019). Pode-se pensar que um modelo *bigram* com 1.000 vocabulários terá 1 milhão de dimensões e um modelo *trigram* terá 1 trilhão de dimensões.

Etapa G: Recuperação dos documentos utilizando a *Application Programming Interface – API* da ferramenta *Apache Solr*.[3] Os resultados da busca são armazenados em arquivo formato *JavaScript Object Notation* (JSON). Para a utilização desta ferramenta alguns passos foram seguidos: a) criação do *core*, b) Definição do *tokenizer*, c) Definição dos filtros (*LowerCase, ASCIIFolding, stopwords, BrazilianSteam*), d) Indexação e simulações na plataforma web do Solr.

Etapa H: Os *datasets* (IZO *et al.*, 2023) já estavam com as entidades anotadas. Desta forma, foram analisados e pré-processados para ficar no padrão da ferramenta de busca. As sentenças foram separadas em linhas através de uma estrutura em formato *eXtensible Markup Language* (XML) e preparadas de acordo com as *tags* de entrada do *Apache Solr*.

Etapa I: Ranqueamento dos documentos mais relevantes, utilizando o algoritmo de pontuação conhecido como modelo *Term Frequency - Inverse Document Frequency* (TF-IDF). As pontuações são normalizadas para que fiquem entre 0 e 1. Os documentos listados em ordem decrescente, sendo as pontuações mais altas consideradas os resultados mais precisos.

Etapa J: Visão do usuário ao visualizar os resultados através de uma plataforma *web*. Esta funcionalidade foi desenvolvida com tecnologia de HTML5 juntamente com a linguagem *Python*.

4. Resultados e Experimentos

Os experimentos desempenham papel importante para uma pesquisa, pois permitem a análise e validação dos resultados de uma investigação científica. Esta Seção descreve como foi a condução dos experimentos, detalhando o planejamento, os resultados e discussões e por fim relatando as limitações deste projeto.

4.1. Planejamento dos experimentos

Para avaliar a abordagem proposta neste trabalho, foram selecioandos aleatoriamente 15 planos de ensino disponíveis na *Web* em instituições de ensino. Estes planos de ensino continham informações básicas como o nome da instituição e do docente,

[3] ⟨https://solr.apache.org/⟩

nome do componente curricular, ementa, objetivos geral e específico, metodologia, conteúdo programático, avaliações e referências. Estes planos de ensino serviram de entrada para que o sistema pudesse obter os materiais complementares. Esta etapa está representada na Figura 1-B.

A Tabela 1 descreve algumas informações sobre os planos de ensino selecionados. Para facilitar a explicação e posterior discussão sobre os resultados, os experimentos foram denominados como: PE01 referente ao Plano de Ensino 1, PE02 referente ao Plano de Ensino 2 e assim sucessivamente. Omitimos informações pessoais para não expor os autores dos documentos e a instituição de ensino.

Tabela 1. Planos de ensino selecionados e utilizados nos experimentos

Experimento	**Componente Curricular**	**CH**
PE01	Química Orgânica e Biológica	108
PE02	Análise Orgânica	108
PE03	Química Orgânica II	36
PE04	Química Orgânica Experimental	72
PE05	Química Orgânica Biológica Teórica	54
PE06	Fundamentos da Química Geral e Orgânica	72
PE07	Química Orgânica	72
PE08	Química Orgânica II	72
PE09	Química Orgânica I	36
PE10 e PE15	Análise Orgânica	36
PE11	Química Orgânica Teórica B	72
PE12	Química Orgânica Teórica C	72
PE13	Química Orgânica Teórica A	72
PE14	Química Orgânica I	96

O conjunto de dados que serviu para recuperar os documentos foi obtido em Izo *et al.* (2023). Os dados estão em formato *eXtensible Markup Language (XML)* e foram extraídos dos arquviso originais em formato *Portable Document Format* (PDF) disponíveis nos sites dos Instituto Nacional de Propriedade Industrial (INPI) e da Revista Virtual de Química (RVQ). Assim, o *dataset* é composto por 30 patentes químicas e 20 artigos científicos. A Figura 1-H representa a etapa em que os *datasets* são consultados para processamento e recuperação de informações para gerar os materiais complementares.

O conjunto de dados de artigos científicos e patentes contém as seguintes categorias de entidades químicas reconhecidas: *Classes Químicas, Compostos Químicos, Elementos Químicos, Equipamentos Químicos e Métodos Químicos*. A Tabela 2 apresenta a quantidade de entidades anotadas para cada tipo de categoria química. Com os dados selecionados, iniciou-se a fase de experimentos, que está descrita à seguir.

Tabela 2. Quantidade de Entidades Anotadas (IZO *et al.*, 2023)

Categoria	***Elementos***	***Compostos***	***Classes***	***Equipamentos***	***Métodos***	**Total**
datasetPat	1422	614	378	745	553	3712
datasetArt	417	65	43	104	441	1070

4.2. Resultados

A fase de experimentos considerou cada plano de ensino como parte de um experimento individual. Assim, realizou-se 15 experimentos seguindo o fluxo

apresentado na Figura 1.

A Tabela 3 apresenta a distribuição do resultado dos experimentos. Nesta tabela foram identificada as informações como siglas para facilitar a representação da tabela.

Tabela 3. Resultado dos experimentos

NExp	QERPE	QCREN	QTIPE	QCR	QASMC	QPSMC	PCR
PE01	4	17	82	21	40	40	25.60
PE02	3	7	24	21	33	25	87.50
PE03	20	63	46	22	20	43	47.82
PE04	4	28	12	8	18	18	66.70
PE05	4	17	96	39	59	51	40.62
PE06	23	82	170	62	86	101	36.47
PE07	8	19	40	23	42	56	57.50
PE08	6	17	44	22	49	65	50.00
PE09	7	33	54	21	33	40	38.88
PE10	13	79	90	41	93	81	45.55
PE11	29	86	100	45	62	86	45.00
PE12	11	31	104	34	36	46	32.69
PE13	4	20	104	41	42	48	39.42
PE14	4	31	104	34	33	26	32.69
PE15	5	24	38	19	33	36	50.00
Total	145	554	1.108	453	679	762	46.43

***Número do Experimento (NExp)*:** É referente ao número do experimento. As siglas PE01, PE02, PE03 [...] PE15 são abreviações para os planos de ensino selecionados.

***Quantidade de Entidades Reconhecidas nos Planos de Ensino (QERPE)*:** É descrita a quantidade de entidades nomeadas reconhecidas nos planos de ensino após a aplicação do algoritmo. Foram encontradas ao todo 145 entidades nos arquivos de planos de ensino. Vale ressaltar que a utilização de buscadores convencionais, tais como *Google*, *Bing*, *Yahoo* e outros, fazem buscas baseando-se em *N-Grams*. Os autores Izo *et al.* (2023) mostram que buscas fundamentadas em estruturas semânticas mais ricas de significados são mais eficientes. Os resultados a seguir fazem uso de uma abordagem híbrida entre a proposta clássica e aquela proposta no artigo previamente citado.

***Quantidade de Conteúdos Recuperados a partir das Entidades Nomeadas (QCREN)*:** Esta coluna se refere a quantidade de tópicos (conteúdos) que o algoritmo recuperou a partir da pesquisa nas bases de dados de *patentes* e *artigos científicos* utilizando como parâmetro de busca as entidades nomeadas reconhecidas nos planos de ensino. Foram recuperados 554 documentos entre artigos científicos e patentes. O quantitativo é maior do que o número de documentos dos conjuntos de dados porque um mesmo artigo ou patente foi sugerido mais de uma vez para uma das entidades nomeadas pesquisadas.

***Quantidade de Tópicos Identificados nos Planos de Ensino (QTIPE)*:** Além das entidades nomeadas, aplicou-se técnicas de *N-Gram* conforme explicado na *etapa F* da Seção 3. Utilizou-se o modelo *2-Gram* pois este apresentou bons resultados sem diminuir tanto o desempenho do algoritmo. Identificou-se que os planos de ensino possuem uma estrutura onde, na maioria das vezes, os docentes listam tópicos com uma única palavra. Assim, identificou-se que alguns conteúdos, por exemplo *haletos* e *alquile*, haviam sido desconsiderados ao aplicar o padrão *2-Gram*. Desta forma, complementou-se a seleção

dos conteúdos com a utilização do padrão *1-Gram* nas senteças de palavras únicas, após a retirada dos *stopwords*. Assim o conjunto de tópicos identificados serviu de base para a pesquisa nos conjuntos de dados de *patentes* e *artigos científicos*.

***Quantidade de Conteúdos Recuperados (QCR)*:** Este item se refere a quantidade de tópicos (conteúdos) que o algoritmo recuperou a partir da pesquisa nas bases de dados de *patentes* e *artigos científicos*. Esses dados servem como parâmetro para avaliar a recuperação das informações.

***Quantidade de Artigos Sugeridos como Material Complementar (QASMC)*:** Quantidade de artigos sugeridos como material de apoio após a consulta na base de artigos científicos. É importante citar que na maioria das vezes o valor é mais alto que a quantidade de artigos científicos da base. Isso acontece porque um mesmo artigo foi sugerido mais de uma vez para um dos conteúdos citados no *QTIPE*.

***Quantidade de Patentes Sugeridas como Material Complementar (QPSMC)*:** Quantidade de patentes sugeridas como material de apoio após a consulta na base de patentes. Conforme aconteceu na seleção de artigos, o quantitativo de artigos é mais alto que a quantidade de patentes da base porque uma mesma patente foi sugerida mais de uma vez para um dos conteúdos citados no *QTIPE*.

***Porcentagem de Conteúdos Recuperados (PCR)*:** Aqui é a porcentagem de conteúdo recuperados em relação ao quantitativo de conteúdo selecionados para a pesquisa. O cálculo se baseia na relação entre *QTIPE* e *QCR*.

A partir dos resultados obtidos é importante fazer algumas ponderações. Quanto a análise das bases, observou-se que a base de patentes conseguiu recuperar mais resultados do que a base de artigos. Isso se deve ao fato de disponibilizarmos os testes com mais patentes em relação ao número de artigos científicos. Por outro lado, os artigos tinham mais textos do que as patentes, haja vista que patentes possuem muitas imagens e tabelas.

Somente com a utilização de entidades nomeadas foram recuperadas cerca de 554 referências existentes dentre os 50 documentos utilizados pelo buscador. Este número corresponde a quase 40% do total de referências a documentos conseguidos quando utilizado o padrão *2-Gram* (QASMC + QPSMC). É importante citar que todas as entidades nomeadas conseguiram recuperar a indicação de pelo menos um documento como material complementar.

Ao analisar os planos de ensino, percebeu-se que quanto melhor descrito estiver o conteúdo programático, melhores serão os resultados. Verificou-se que os planos de ensino PE05 e PE06 e os planos entre PE10 e PE14 foram os que possuiam os conteúdos mais bem descritos e isso refletiu na relação maior de tópicos e entidades nomeadas reconhecidas.

O PE02 foi o que teve os melhores resultados. Este plano foi analisado e observou-se que o conteúdo era extenso e com termos técnicos, além de palavras compostas que permitiram o reconhecimento pelo padrão *2-Gram*. Desta forma, o algoritmo conseguiu efetuar buscas mais precisas nas bases de dados. O pior valor foi encontrado no PE01. Este plano de ensino foi o menor plano utilizado nos experimentos. Essa informação corrobora com a análise de quanto melhor descrito o plano, melhor são as alternativa para recuperar os conteúdos complementares.

4.3. Limitações

Esta pesquisa apresenta algumas limitações. A mais significativa é a necessidade de anotação das bases de dados consultadas, a fim de aprimorar os resultados. Um

ponto positivo é que as bases que já estão anotadas servem de treino para uma anotação automáticas das novas bases.

A LG*q* Plus utilizada nesta pesquisa ainda precisa de melhorias, pois grande parte das anotações são referentes a área de química orgânica. Essa análise se refere principalmente às anotações feitas nas categorias *Compostos químicos* e *Classes químicas*. Por esta razão, os planos de ensino selecionados são, em sua maioria, sobre a área *química orgânica*.

5. Conclusões e Trabalhos Futuros

A proposta apresentada neste trabalho utiliza-se de uma bordagem híbrida onde a extração e utilização de entidades nomeadas bem como *N-Grams* encontrados em documentos de plano de ensino nos permitiu realizar um processo de recuperação de informação em artigos científicos e patentes de forma bem mais eficiente. De fato, como já demonstrado em Izo *et al.* (2023), essa forma mais rica da representação semântica de estruturas existentes nos textos nos permitiu aumentar as métricas de recuperação dos documentos de interesse.

A anotação de entidades é parte importante do processo para que se consiga gerar informações mais completas e próximas do objetivo do usuário (PIROVANI; OLIVEIRA, 2021). Assim, é importante que os documentos estejam anotados para que sirvam de pesquisa pelos planos de ensino. Se novos documentos (artigos e patentes) forem agregados ao sistema, não será necessário anotar tudo manualmente, pois poderemos utilizar os documentos já anotados e validados para treinar e fazer anotações nos novos documentos.

Em trabalhos futuros além de investigar alterantivas para melhorar os resultados das buscas, há a intenção de investigar também os planos de aula para analisar conteúdos mais específicos. Também planeja-se expandir o conjunto de dados, pois essa abordagem pode contribuir para melhorar os resultados, uma vez que teremos uma maior quantidade de materiais complementares disponíveis para anotação.

Referências

ABURAJAB, A.; SALMAN, A. Interactive Blackboard for Web-based Real-time Tutoring System. In: **2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)**. Amman, Jordan: IEEE, 2019. p. 63–68.

AHMAD, A.; YAACOB, N. N. M. Embedding Teaching Plan into E-learning System. In: YACOB, N. A.; NOOR, N. A. M.; YUNUS, N. Y. M.; YUSSOF, R. L.; ZAKARIA, S. A. K. Y. (Ed.). **Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016)**. Singapore: Springer Singapore, 2018. p. 89–96.

ALFERES, M. A.; MAINARDES, J. Um currículo Nacional para os Anos Iniciais? Analise Preliminar do Documento. **Currículo sem Fronteiras**, v. 14, n. 1, p. 243–259, 2014.

ASTUTIK, I. R. I.; ROSID, M. A. Integrated Information System Teaching Plan in College Using FAST Method and Twitter Bootstrap. **Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control**, p. 163–170, 2018.

CAN, Y. S.; KABADAY1, M. E. Automatic Estimation of Age Distributions from the First Ottoman Empire Population Register Series by Using Deep Learning. **Electronics**, v. 10, n. 18, 2021.

CERI, S. *et al.* An Introduction to Information Retrieval. In: ____. **Web Information Retrieval**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 3–11.

CONEGLIAN, C. S. **Recuperação da Informação com Abordagem Semântica Utilizando Linguagem Natural: a Inteligência Artificial na Ciência da Informação**. Tese (Doutorado) — Tese (Ciência da Informação-FFC)–Universidade Estadual Paulista UNESP–SP, 2018.

FERREIRA, M.; LOPES, M. **Para Conhecer: linguística computacional**. São Paulo: Editora Contexto, 2019. 192 p.

FERREIRA, M.; REHFELDT, H.; SILVA, J. Tecnologias digitais no ciclo de alfabetizaÇÃo: Analisando um projeto polÍtico pedagÓgico e os planos de ensino de uma professora. **Imagens da Educação**, v. 10, p. 01–15, 03 2020.

GROSS, M. The Construction of Local Grammars. **Finite-state language processing**, v. 1, p. 329, 1997.

HILL, M. J.; HENGCHEN, S. Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study. **Digital Scholarship in the Humanities**, Oxford University Press, v. 34, n. 4, p. 825–843, 2019.

IZO, F.; VEREAU, L. E.; PIROVANI, J.; OLIVEIRA, E.; BADUE, C. An Intelligent Report Generator For Chemical Documents. In: **Anais do XIX Simpósio Brasileiro de Sistemas de Informação**. Porto Alegre, RS, Brasil: SBC, 2023.

KARTHIKEYAN, S.; HERRERA, A. G. S. D.; DOCTOR, F.; MIRZA, A. An OCR Post-Correction Approach Using Deep Learning for Processing Medical Reports. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 32, n. 5, p. 2574–2581, 2021.

MOOERS, C. N. Zatocoding Applied to Mechanical Organization of Knowledge. **American documentation**, Wiley Online Library, v. 2, n. 1, p. 20–32, 1951.

PANDEY, B. K. *et al.* Effective and Secure Transmission of Health Information Using Advanced Morphological Component Analysis and Image Hiding. In: SPRINGER. **Artificial Intelligence on Medical Data: Proceedings of International Symposium, ISCMM 2021**. Singapura, 2022. p. 223–230.

PIROVANI, J.; OLIVEIRA, E. Studying the Adaptation of Portuguese NER for Different Textual Genres. **JofSuper**, Springer International Publishing, p. 1–17, 2021.

SARACEVIC, T. Information Science: Origin, Evolution and Relations. **Perspectivas em Ciência da Informação; v. 1, n. 1 (1996)**, v. 24, n. 2, 1992.

SOPER, E.; FUJIMOTO, S.; YU, Y.-Y. BART for Post-Correction of OCR Newspaper Text. In: **Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)**. Online: Association for Computational Linguistics, 2021. p. 284–290.

WU, Z.; SHEN, S.; LI, H.; ZHOU, H.; LU, C. A Basic Framework for Privacy Protection in Personalized Information Retrieval: An Effective Framework for User Privacy Protection. **Journal of Organizational and End User Computing (JOEUC)**, IGI Global, v. 33, n. 6, p. 1–26, 2021.

ZHANG, L.; HOU, W. An Automatic Construction Technology of Chinese Knowledge Graph for Teaching Plan. In: SUN, X.; ZHANG, X.; XIA, Z.; BERTINO, E. (Ed.). **Advances in Artificial Intelligence and Security**. Cham: Springer International Publishing, 2022. p. 95–105.