UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Sistema de recomendação de agentes de controle biológico contra patógenos de plantas baseado em grafos de conhecimento

Francisco Marques Vicente

Mestrado em Engenharia Informática

Dissertação orientada por:
Prof.ª Doutora Márcia Cristina Afonso Barros
Doutor Ricardo Filipe Serrote Ramiro

2024

Agradecimentos

Em primeiro lugar, gostaria de expressar o meu profundo agradecimento à minha orientadora, Prof. Márcia Barros, pelo apoio e orientação incansável ao longo de todo este percurso, fundamentais para a concretização deste projeto. Ao meu coorientador, Dr. Ricardo Ramiro, agradeço pela generosa partilha do seu conhecimento na área da biologia, que tanto contribuiu para o desenvolvimento deste trabalho.

Agradeço igualmente ao meu colega e amigo João, pela sua constante disponibilidade para discutir ideias e pelo apoio incondicional ao longo de todo o projeto. Um agradecimento especial à Catarina pela sua prontidão em colaborar e ajudar sempre que necessário.

Dedico esta tese à minha família, pelo apoio incondicional e por me proporcionarem a possibilidade de seguir este caminho.

Por fim, quero expressar a minha gratidão à Beatriz, pela sua motivação, amizade e amor, que foram uma força constante ao longo desta jornada.

Resumo

Perante crescentes preocupações com os riscos ambientais e para a saúde associados aos pesticidas químicos, a necessidade de meios mais sustentáveis no combate contra patógenos nas plantas aumenta. A descentralização da informação sobre agentes de controlo biológico, leva a que a descoberta de novos micróbios capazes de inibir tais patógenos e a sua aplicabilidade nos campos seja bastante demorada e dispendiosa.

De forma a combater esse problema, este trabalho foca-se em aproveitar o poder dos grafos de conhecimento e da aprendizagem automática, para reunir todas as associações conhecidas entre micróbios e patógenos, criando um sistema de recomendação de novos micróbios antagonistas para determinados patógenos de plantas.

Com base na metodologia LIBRETTI, utilizámos o MERpy para identificar micróbios presentes na literatura científica sobre patógenos, extraída do PubMed, e criar o conjunto de dados PPathoRM no formato <utilizador, item, *rating*>, sendo os utilizadores os patógenos, os itens os micróbios e o *rating* o número de artigos que aparecem em comum. Aplicamos o algoritmo KNN e ALS, concluindo que o ALS apresenta os melhores resultados de recomendação independente do número de itens recomendados para PPathoRM.

Além disso extraímos o tipo de relação presentes nos artigos científicos entre micróbios e patógenos para integrar no grafo do conhecimento PPathoKG. A comparação da avaliação dos resultados, algoritmo de embeddings de grafo de conhecimento DistMult no PPathoKG com o algoritmo ALS em PPathoRM, demonstra que a integração de relações e o seu uso na recomendação melhora a qualidade do sistema.

Palavras-chave: Agentes de Controlo Biológico, Grafo de Conhecimento, Sistema de Recomendação, Extração de Relações, Reconhecimento de Entidades Nomeadas

Abstract

In light of growing concerns about the environmental and health risks associated with chemical pesticides, the need for safer methods to combat plant pathogens is becoming more urgent. The decentralization of information on biological control agents makes the discovery of new microbes capable of inhibiting these pathogens, as well as their practical application in the field, a slow and costly process.

To address this issue, this work focuses on leveraging the power of knowledge graphs and machine learning to consolidate all known associations between microbes and pathogens, thereby creating a recommendation system for new antagonistic microbes to target specific plant pathogens.

Using the LIBRETTI methodology, we employed MERpy to identify microbes found in scientific literature related to pathogens, extracted from PubMed, and created the PPathoRM dataset in the <user, item, rating>format, where the users are pathogens, the items are microbes, and the rating reflects the number of articles in which both appear together. We applied the KNN and ALS algorithms, concluding that ALS delivers the best recommendation results, regardless of the number of items recommended for PPathoRM.

Moreover, we extracted the types of relationships between microbes and pathogens described in scientific articles to incorporate them into the knowledge graph PPathoKG. Comparing the results obtained from applying the DistMult knowledge graph embedding algorithm in PPathoKG with the ALS algorithm in PPathoRM shows that integrating relationships and using them in recommendations significantly enhances the system's performance.

Keywords: Biological Control Agents, Knowledge Graph, Recommendation System, Relation Extraction, Named Entity Recongnition

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A produção de alimentos tem-se tornado uma das maiores preocupações a nível global, devido ao aumento da população e às alterações climáticas nos anos mais recentes, impactando diretamente o aparecimento de novos patógenos nas plantas, a regularidade com que estes começam a aparecer e como se tornam cada vez mais capazes de se espalhar a outras regiões, afetando novas plantas [77].

O grupo funcional dos fitopatógenos inclui diversos organismos como vírus, bactérias e fungos que constituem uma ameaça considerável para a saúde das plantas. O resultado das suas funções biológicas nomeadamente nutrição, desenvolvimento e reprodução desencadeia um impacto negativo a nível celular e sistemático do seu hospedeiro [9]. A sua ação pode ser direta, causando a necrose celular ou prolongada, mantendo uma relação biotrófica com o hospedeiro de forma a assegurar a sua sobrevivência a custo da produtividade da planta [61].

As doenças causadas por bactérias, fungos ou vírus, são uma das principais causas da insegurança alimentar e da perda de grandes quantidades da colheita [77]. Este cenário exige a procura constante de métodos eficazes para a proteção das culturas e a garantia da segurança alimentar a longo prazo.

Alguns avanços têm sido feitos na direção de melhorar as práticas agrícolas, facilitando o trabalho dos agricultores na melhoria e manutenção das suas culturas. Novas ferramentas baseadas em inteligência artificial permitem poupar tempo e reduzir os custos relacionados com a deteção, classificação e tratamento de doenças nas plantas [42, 4]. Essas inovações superam os métodos antiquados dependentes da experiência e do conhecimento do agricultor ou de profissionais contratados, baseados em avaliações manuais, proporcionando uma abordagem mais precisa e eficiente [79]. Os sistemas de recomendação, como parte dessas novas ferramentas, facilitam a tomada de decisão dos agricultores ao complementar a informação adquirida, oferecendo recomendações personalizadas adaptadas às suas necessidades [62].

1.1 Motivação

Para o tratamento destas doenças, o uso de pesticidas químicos tem sido o meio de controlo mais utilizado, devido à sua eficácia e rapidez no combate ou inibição dos patógenos. No entanto,

estes pesticidas têm, por vezes, um impacto negativo na saúde humana e ambiental [47], além de contribuírem para o desenvolvimento de resistências nos patógenos [55]. Estas questões levantam a necessidade urgente de encontrar alternativas sustentáveis que protejam as culturas sem causar danos colaterais.

Reconhecendo que a produção agrícola necessita de se tornar mais sustentável e promover a segurança alimentar, a União Europeia introduziu recentemente a estratégia ”Farm to Fork”como parte da sua iniciativa do ”European Green Deal”[96]. Esta estratégia procura melhorar as práticas agrícolas, reduzindo o uso de pesticidas químicos e identificado as soluções biológicas como possíveis alternativas à utilização de pesticidas químicos. Entre estas soluções encontram-se os agentes de controlo biológico, que são microrganismos capazes de matar ou inibir a multiplicação dos agentes patogénicos, sem causar efeitos negativos nas plantas.

Atualmente, diversas empresas biotecnológicas, alinham com as diretrizes do ”European Green Deal”, dedicando-se à procura e fabricação destes tipos de produtos mais biológicos. No entanto, a sua capacidade de desenvolver novos produtos biológicos é limitada pela falta de ferramentas adequadas para apoiar a descoberta de microrganismos eficazes. A procura por micróbios com potencial antagonista é particularmente desafiante devido à escassez de ferramentas especializadas que facilitem o processo de identificação, seleção e validação destes agentes. Devido à crescente necessidade de reduzir o impacto ambiental da agricultura, é necessário que sejam desenvolvidas ferramentas inovadoras que acelerarem a descoberta e produção de produtos compostos por agentes de controlo biológicos.

1.2 Problema

Apesar dos microrganismos serem uma opção potencialmente mais benéfica que as soluções químicas, ainda são pouco utilizados na agricultura. O processo de identificação de microrganismos eficazes é bastante demorado e dispendioso para as empresas biotecnológicas que necessitam de os descobrir. Devido a esta dificuldade, o desenvolvimento de novos agentes de controlo biológico tem-se focado principalmente num conjunto restrito de espécies, pertencentes aos géneros *Bacillus*, *Pseudomonas* e *Trichoderma*, que obedecem aos seguintes critérios, cultiváveis no laboratório, alta probabilidade de uma estirpe destas espécies ter potencial de agente de controlo biológico e mecanismos de ação bem estudados [96]. No entanto, existem milhares de espécies de microrganismos associados com as plantas, muitas das quais poderão ter potencial como agentes de controlo biológico. Identificar essas espécies permite alargar a diversidade de modos de ação dos agentes de controlo biológico, mas este processo é dificultado pela falta de informação estruturada sobre os mesmos. Por outro lado, relativamente aos produtos já utilizados no campo, verifica-se que os agricultores têm frequentemente dúvidas acerca de quais os produtos mais indicados para controlar determinada doença. Como tal, a existência de sistemas de recomendação de agentes de controlo biológico facilitaria a própria atividade agrícola.

Os sistemas de recomendação, que são mecanismos baseados em algoritmos de análise de dados, podem fornecer novas recomendações personalizadas de acordo com os comportamentos de

um utilizador [46]. No contexto da agricultura, estes sistemas podem desempenhar um papel crucial na identificação eficiente de microrganismos benéficos, capazes de combater certos patógenos nas plantas. No entanto, a existência de ferramentas de aconselhamento específicas para micróbios como agentes de controlo biológico contra patógenos ainda é pouco explorada.

Além disso, a falta de organização e estruturação dos dados sobre as relações e interações entre patógenos de plantas e micróbios antagonistas torna difícil tanto encontrar bases de dados adequadas como desenvolver novas fontes de informação acessíveis e eficientes. Esta informação fica muitas vezes dispersa e perdida em grandes repositórios de artigos científicos, como por exemplo o PubMed[1].

1.3 Objectivos

O objetivo desta tese é desenvolver um grafo de conhecimento que contenha relações patógeno-micróbio, de forma a criar um sistema de recomendação que indique micróbios prováveis de serem capazes de inibir certos tipos de patógenos. Este sistema de recomendação facilitará a identificação dos microrganismos mais eficazes para a proteção das culturas no combate a diferentes doenças de plantas. Com esta abordagem, pretende-se contribuir para uma agricultura mais sustentável e eficiente, alinhada com as diretrizes do "European Green Deal".

Este trabalho divide-se em 4 etapas diferentes:

- Recolha e tratamento dos dados de diferentes origens;
- Criação do conjunto de dados com as associações entre patógenos e micróbios;
- Construção do grafo de conhecimento;
- Desenvolvimento do sistema de recomendação.

1.4 Contribuições

As contribuições do trabalho desenvolvido são:

- Um conjunto de dados específico com diferentes associações entre patógenos de plantas e micróbios;
- Um grafo de conhecimento com relações entre patógenos de plantas e micróbios;
- Um sistema de recomendação de micróbios com potencial para inibir patógenos de plantas, disponível em `https://github.com/franciscomvicente/PlantPathoRec`.
- 2.º Melhor Poster do 9.º LASIGE Workshop.

[1] https://pubmed.ncbi.nlm.nih.gov/

1.5 Estrutura do Documento

Este documento está organizado da seguinte forma:

- Capítulo 1: Introdução, apresenta a motivação, problema objetivos, metodologia, contribuições e a organização do documento;
- Capítulo 2: Enquadramento Teórico, apresenta todos os conceitos necessários para perceber o trabalho;
- Capítulo 3: Trabalho Relacionado, apresenta toda a pesquisa realizada antes do trabalho, quais os algoritmos e ferramentas existentes no seu estado de arte;
- Capítulo 4: Metodologia, descreve todas os passos na resolução deste trabalho, referindo todas as ferramentas e algoritmos usados;
- Capítulo 5: Resultados, apresenta uma análise detalhada do conjunto de dados, do grafo de conhecimento e das ferramentas utilizadas, validando a qualidade do conjunto de dados e do grafo de conhecimento para um sistema de recomendação.
- Capítulo 6: Conclusão, resume os principais resultados alcançados, discute as limitações do trabalho realizado, sugere possíveis direções para trabalhos futuros e reflete sobre o impacto e a relevância das contribuições para a área de estudo.

Capítulo 2

Enquadramento Teórico

Neste capítulo serão apresentados todos os conceitos chave para compreender o problema e as ferramentas usadas.

2.1 Micróbios e Patógenos

Micróbios, ou microrganismos, são organismos microscópicos invisíveis a olho nu que constituem as unidades fundamentais das complexas comunidades conhecidas como microbiota, compostas por milhares de indivíduos [34]. Essas comunidades microbianas existem numa ampla variedade de ambientes naturais, manifestando uma diversidade de interações biológicas, tais como competição, parasitismo, mutualismo e cooperação. Desempenham importantes funções mantendo o equilíbrio ecológico, assegurando a produtividade agrícola e a segurança alimentar [27].

Nesse ecossistema, a presença de patógenos adiciona uma camada de complexidade. Os patógenos, microrganismos prejudiciais que são capazes de causar doenças, interagem ativa e continuamente com outros elementos da microbiota e com os seus hospedeiros. Essas interações resultam em processos de competição pelos mesmos recursos, assim como no desenvolvimento de novas estratégias de resistência por parte dos hospedeiros [15, 16].

O não controlo destes patógenos compromete a saúde e a produtividade dos hospedeiros, pelo que a interação entre micróbios e patógenos é fulcral, pois além das suas funções principais, alguns destes micróbios são capazes de inibir o crescimento e propagação dos patógenos, servindo de controlo biológico para pragas e doenças das plantas [27].

2.2 Grafos de Conhecimento

Os Grafos de Conhecimento são bases de conhecimento estruturadas em grafos que desempenham um papel crucial na representação e na obtenção de informação. Estes são capazes de representar informações do mundo real, sendo uma ferramenta poderosa para modelar entidades e as suas relações, tornando-os bastante bem reconhecidos para representar informação complexa, facilitando a compreensão da informação pelas máquinas [92]. Este conceito tem uma longa história nos campos da lógica e da inteligência artificial e tem sido frequentemente referido como “redes

semânticas", pelo seu foco em capturar e estruturar o significado das relações entre conceitos.

O termo de grafo de conhecimento começou a ganhar destaque a partir de 2012, quando a Google introduziu a sua estrutura de bases de conhecimento baseada em grafos contendo milhões de entidades e as suas relações associadas [26]. A sua aplicabilidade nos motores de pesquisa da empresa teve uma enorme visibilidade e relevância, permitindo obter respostas mais exatas e precisas para as questões dos utilizadores.

A sua utilidade tem vindo a aumentar podendo diferenciar este em dois tipos diferentes: grafos de conhecimento abertos, como DBpedia [49] e Wikidata [89], ou grafos de conhecimento empresariais. A sua aplicabilidade em vários tipos de áreas é cada vez mais comum [66], devido à sua versatilidade e poder de modelação dos dados.

Um grafo é definido por um conjunto de vértices e um conjunto de arestas, organizados em um formato de triplo. Cada triplo contém dois nós, que representam duas entidades distintas que têm uma ligação entre elas, e por um vértice, representando a ligação entre estas duas entidades. Enquanto um grafo se concentra na conexão e estrutura das ligações entre os nós, sem atribuir um significado intrínseco a essas ligações, o grafo de conhecimento distingue-se pela sua capacidade em representar informações do mundo real. Nos grafos de conhecimento, cada ligação representa um tipo de relação, tendo um significado específico, permitindo a distinção de diferentes interações entre entidades. Além disso, estes permitem a inferência de novas informações a partir das relações existentes.

A Figura 2.1 ilustra um exemplo de um grafo do conhecimento, onde é possível ver, por exemplo, que os nós *e1* e *e2* estão conectados pela aresta *r1*, indo de *e2* para *e1* formando um triplo (*e2*,*r1*,*e1*). A aresta *r1* representa a relação existente entre a entidade *e2*

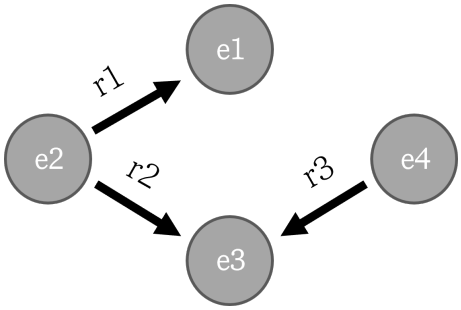

Figura 2.1: Exemplo genérico de um grafo de conhecimento

O armazenamento de grafos de conhecimento é geralmente feito em bases de dados de grafos. Neo4j [1] é uma base de dados de grafos NoSQL, popular e amplamente utilizada. É disponibilizada em código aberto, com suporte total e uma opção comercial, possuindo um esquema flexível e escalável. Esta utiliza Cypher, uma linguagem de consulta declarativa projetada especificamente para bancos de dados de grafos [33]. Destaca-se pela sua capacidade de modelação e visualização do grafo de maneira intuitiva, permitindo uma interação com os dados mais compreensível e deta-

[1] https://neo4j.com/

lhada. Existem ainda outras como por exemplo TigerGraph [2], JanusGraph [3] e ArangoDB [4].

2.3 Processamento de Linguagem Natural

O Processamento de Linguagem Natural é uma sub-área da inteligência artificial, que se foca na interação dos computadores com a linguagem humana [11]. Esta utiliza métodos e algoritmos para representar sintaticamente o texto, permitindo que as máquinas sejam capazes de processar, interpretar e produzir a linguagem escrita ou falada de um ser humano, facilitando a sua forma de comunicação [11]. É utilizada em diversas tarefas como análise de sentimentos [43], reconhecimento de voz [21], tradução de texto [2] e extração de informação [78].

2.3.1 Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas é uma importante tarefa no Processamento de Linguagem Natural, que consiste na identificação, localização e classificação de entidades específicas presentes em dados não estruturados, como textos. Estas entidades nomeadas são frases ou palavras que representam uma determinada categoria, como uma organização, uma data, um lugar, ou qualquer outro tipo de categoria predefinida [64].

O Reconhecimento de Entidades Nomeadas pode ser aplicada em diferentes áreas, variando também o tipo de entidades que se pretendem identificar [67]. Para isso existem três abordagens diferentes para criar um modelo de Reconhecimento de Entidades Nomeadas. Modelos baseado em regras, que utilizam regras ou padrões predefinidos, analisando a estrutura e gramática das frases. Modelos baseados em dicionários, utilizando lista pré-definidas de entidades, que se pretendem identificar nos textos. Modelos baseados em aprendizagem automática, que usam modelos supervisionados treinados com diferentes conjuntos de dados anotados e que permitem generalizar para diferentes domínios.

A Figura 2.2 representa o funcionamento do uso desta ferramenta num texto, realçando as entidades encontradas. É possível ver que esta identificou a entidade '25 de Abril de 1974' com a categoria Data, as entidades 'Movimento das Forças Armadas' e 'Estado Novo' como Organizações, 'Revolução dos Cravos' como Obra e 'Lisboa' como um Local.

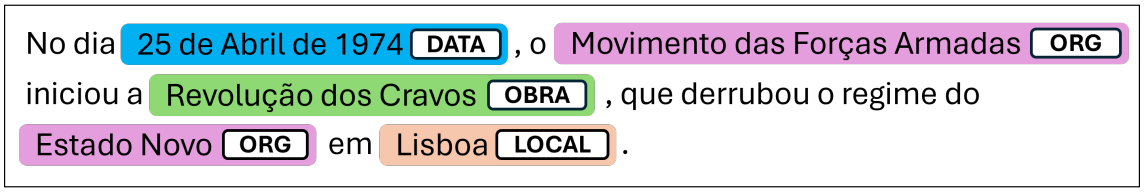

Figura 2.2: Exemplo de Reconhecimento de Entidades Nomeadas num texto, destacando as entidades identificadas, assim como a sua categoria.

[2] https://www.tigergraph.com/
[3] https://janusgraph.org/
[4] https://arangodb.com/

2.3.2 Extração de Relação

A Extração de Relações, tal como o reconhecimento de entidades nomeadas, é uma tarefa essencial no Processamento de Linguagem Natural cujo objetivo é identificar e classificar automaticamente as ligações entre diferentes entidades mencionadas num texto [64]. Esta foca-se em entender como essas entidades estão relacionadas, analisando o contexto e as palavras que as rodeiam, tentando compreender que tipo de relação está presente, extraindo conhecimento de dados não estruturados e guardando geralmente em bases de dados de conhecimento.

A frase ”Several strains of Bacillus subtilis, isolated and characterized in this work, exhibited growth inhibition against B. cinerea of more than 40% in in vitro cultures.”extraída do resumo do artigo Bolivar-Anillo et al. (2021) [12], contém duas entidades nomeadas, Bacillus subtilis e B. cinerea (que se refere a Botrytis cinerea). Na aplicação da extração de relações, essas duas entidades seriam dadas ao modelo, que identificaria a relação de ”inibição”presente no texto, resultando num triplo (Bacillus subtilis,inibição,Botrytis cinerea).

Os Modelos de Linguagem de Grande Escala têm transformado a extração de relações devido à sua capacidade de captar pormenores contextuais de forma mais eficaz do que abordagens tradicionais, como os modelos baseados em regras e os modelos de aprendizagem automática supervisionados [90]. Estes modelos, como BERT ou GPT, são treinados em vastos conjuntos de dados, permitindo-lhes compreender não apenas o significado literal das palavras, mas também a semântica envolvida nas relações entre entidades. Ao serem treinados com textos que contêm múltiplas entidades, os Modelos de Linguagem de Grande Escala são capazes de inferir ligações subjacentes, mesmo que estas não sejam explicitas no texto [91].

2.4 Sistemas de Recomendação

Com o crescimento exponencial da informação disponibilizada mundialmente pela Internet, torna-se complicado para os utilizadores encontrarem as informações que necessitam, obrigando à procura e leitura intensiva de vários conteúdos. Os sistemas de recomendação foram desenvolvidos de forma a combater esse problema, facultando aos utilizadores informações mais precisas de conteúdos que vão de encontro aos interesses destes [95].

A recomendação é feita estimando o *rating* que cada utilizador dará a um certo item, no estilo utilizador-item-*rating*, determinando assim se o utilizador tem um maior ou menor interesse nesse item, resultando numa lista com os itens com os *ratings* mais elevados entre todos. Os itens no topo da lista serão apresentados ao utilizador.

Existem várias abordagens para estimar o *rating* que cada utilizador daria a um item, sendo essas categorias filtragem colaborativa, filtragem baseada em conteúdo e filtragem híbrida [46, 103].

2.4.1 Abordagens de Recomendação

2.4.1.1 Filtragem Colaborativa

Filtragem colaborativa (Figura 2.3) é uma técnica vastamente utilizada em sistemas de recomendação, que se baseia na ideia de que dois utilizadores com preferências e comportamentos semelhantes, vão gostar dos mesmos itens. Esta abordagem depende da similaridade entre utilizadores ou itens para fazer previsões. Esta é dividida em dois métodos distintos baseados em memória e baseados em modelos [46].

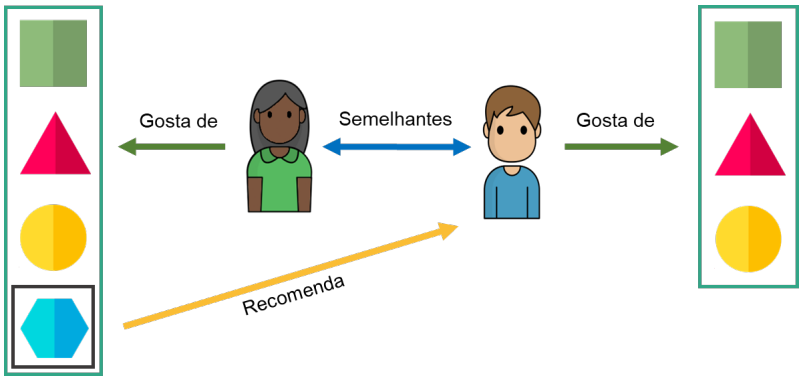

Figura 2.3: Representação do funcionamento da filtragem colaborativa

Os métodos baseados em memória identificam grupos de utilizadores (vizinhança) em que as preferências vão de encontro ao utilizador alvo, sendo as recomendações depois feitas com base nessa vizinhança. Esta pode ser feita de duas maneiras, baseado no utilizador, em que o sistema calcula o *rating* de um novo item para um utilizador considerando os *ratings* dados pela vizinhança de utilizadores que já classificaram positivamente esse item, ou baseado num item construindo a vizinhança de itens semelhantes comparativamente a itens já classificados anteriormente [46]. Este método geralmente utiliza métricas de semelhança como a Semelhança Cosseno, Semelhança de Jaccard, Correlação de Pearson e Semelhança pela Diferença Quadrática Média [80, 6].

Nos métodos baseados em modelos são utilizadas várias técnicas de aprendizagem automática, construindo um modelo preditivo para interesses do utilizador [46]. Os modelos mais usados são, por exemplo Fatorização de Matrizes [50, 53], Redes Neurais [53] e Classificadores Bayesianos [58].

A principal vantagem desta técnica é a recomendação de itens de qualquer categoria, independentemente das suas características intrínsecas, baseando-se apenas nas preferências e interações dos utilizadores, ajudando na descoberta de novos interesses. Por outro lado, esta abordagem enfrenta desafios quando há um número limitado de utilizadores, um fenómeno conhecido como *cold start*, e quando há escassez de dados (*ratings*) [95]. O *cold start* acontece quando novos itens ou utilizadores são adicionados ao sistema, mas por serem recentes, não possuem um histórico de interações, limitando o sistema que depende dessas mesmas interações.

2.4.1.2 Filtragem Baseada em Conteúdo

A filtragem baseada em conteúdo (Figura 2.4) parte da ideia que se um utilizador gostou de um certo item, este vai gostar de itens que tenham atributos semelhantes, agrupando-os tendo em conta o seu perfil de item, com base nas características dos próprios. A similaridade entre os itens pode ser calculada a partir de vetores, usando, por exemplo, Semelhança do Cosseno e Distância Euclidiana [106] ou através de métodos de aprendizagem automática, como Agrupamento e Aprendizagem Profunda [31]. Esta não requer informações de outros utilizadores nem dos seus *ratings*, necessitando apenas de conter características sobre os itens, conseguindo recomendar itens que não sejam muito populares ou que sejam novos (sem qualquer *rating*). Por outro lado, não é capaz de recomendar itens diversos fora do perfil do utilizador e pode ser limitado pela quantidade e qualidade de informação que existe sobre os itens [38].

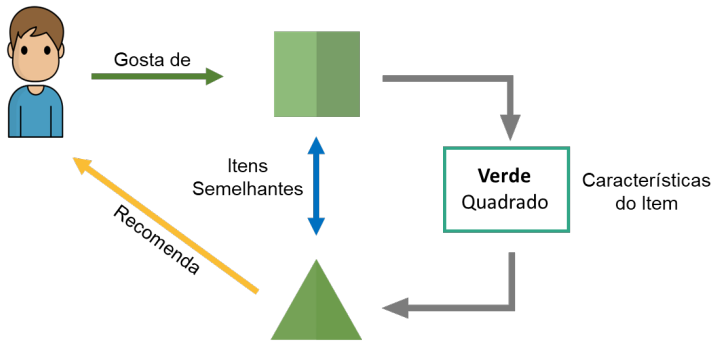

Figura 2.4: Representação do funcionamento da filtragem baseada em conteúdo

2.4.1.3 Filtragem Híbrida

A abordagem híbrida combina dois ou mais métodos juntos para fazer as recomendações, combatendo as limitações das técnicas individuais, melhorando o desempenho e a precisão dos resultados obtidos. O principal desafio é encontrar o conjunto de técnicas combinadas, que produzam o melhor resultado para determinado tipo de problema. Existem vários tipos de métodos que podem ser aplicados para a criação de um sistema híbrido, sendo estes distintos pelo seu design, *ensemble*, *monolithic* ou *mixed* design [3].

O *ensemble* combina o resultado de diferentes algoritmos de forma a obter um resultado mais robusto. Os métodos associados a este design são os *Weighted* que combina os *ratings* de modelos diferentes produzindo um só resultado, os *Switching* que consistem em escolher um algoritmo diferente consoante a situação, e os *Cascade* em que um algoritmo melhora as recomendações feitas por outro algoritmo [14, 37].

O *monolithic* cria um algoritmo integrado usando vários tipos de dados, em que por vezes não dá para distinguir entre filtragem baseada em conteúdo e filtragem colaborativa. Exemplos de métodos são Combinação de Características, onde as caraterísticas provenientes de diferentes fontes de dados de recomendação são agrupadas num único algoritmo de recomendação, Aumento de Características, onde os resultados de uma técnica são usados como características para outra

técnica, e *Meta-level*, onde um modelo criado por um sistema de recomendação é usado por outro sistema recomendação [14, 37].

O *mixed* utiliza as recomendações feitas por diferentes sistemas de recomendação e apresenta-as em conjunto [14, 37].

2.4.2 Tipos de *Feedback*

Os sistema de recomendação baseia-se em interações entre utilizadores e itens para fornecer recomendações personalizadas. A obtenção dessas interações sobre os comportamentos ou preferências de um utilizador pode ser feita de duas maneiras, através de *feedback* explícito ou *feedback* implícito.

O *feedback* explícito refere-se ao input direto dos utilizadores sobre a sua opinião ou preferência perante um determinado item com o qual interagiram. Geralmente, este é recolhido através de *ratings* que um utilizador faz ao avaliar um item, numa escala pré-definida, ou de uma funcionalidade mais simples como de gosto/não gosto. Este tipo de *feedback* reflete diretamente o tipo de interação que cada utilizador tem com cada item, diferenciando entre interações positivas e negativas, além de quantificá-las. No entanto, é limitado por depender na disponibilidade e motivação dos utilizadores em fornecer as suas opiniões, resultando muitas das vezes em uma escassez de dados [40].

Em contraste, o *feedback* implícito não necessita que os utilizadores expressem as suas opiniões sobre um conteúdo. É feita uma recolha passiva dos dados, monitorizando as ações realizadas pelos utilizadores, como os itens que clicam, visualizam, compram, ou o tempo que demoram na sua interação. Este é bastante utilizado quando não existe forma de obter diretamente as preferências do utilizador, ou o número de *ratings* feitos pelos utilizadores é muito reduzido [40]. Contudo, este apenas permite capturar interações positivas. Embora múltiplas interações de um utilizador com um item possam indicar o seu gosto, a ausência de interação não implica necessariamente desinteresse, o que dificulta a interpretação das preferências de forma completa.

2.4.3 Avaliação de Sistemas de Recomendação

É fundamental avaliar o sistema de recomendação para compreender o desempenho, a eficácia e a qualidade das recomendações feitas. Podem ser utilizadas três tipos de abordagens, avaliação online, estudo de utilizador e avaliação *offline*, consoante os requerimentos e do grão de detalhe definidos para avaliação. Na avaliação online o sistema utiliza um ambiente ativo, onde utilizadores reais interagem, normalmente utilizando testes A/B, entendendo como se adaptaria ao mundo real [5]. No estudo de utilizador é simulado um ambiente real, normalmente num laboratório, onde uma pequena amostra de utilizadores interagem com o sistema, permitindo obter resultados e opiniões em tempo real, assim como avaliar diferentes tipos de métricas com o uso de mais sensores. A avaliação *offline* utiliza dados passados de *ratings*, fazendo com que não seja necessário criar um ambiente real para testar o sistema de recomendação, mas limitando a análise. Dependendo do sistema de recomendação, o tipo de avaliação também varia existindo três tipos

de métricas [103].

As métricas de precisão, que são usadas para avaliar algoritmos que preveem o *rating* que um utilizador dá a um determinado item. As mais comuns são Mean Absolute Error (MAE - Equação 2.1), Mean Squared Error (MSE - Equação 2.2) e Root Mean Squared Error (RMSE - Equação 2.3). MAE é a média das diferenças absolutas entre os valores previstos para cada um dos itens e os valores reais dos itens. MSE é a média dos quadrados das diferenças entre os valores previstos para cada um dos itens e os valores reais dos itens. O RMSE é a raiz quadrada do MSE, tornando o valor mais compreensível.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{2.1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{2.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.3}$$

As métricas de classificação, servem para avaliar algoritmos que recomendam uma lista de itens, capturando a proporção de recomendações corretas, sendo as mais comuns Precison (Equação 2.4), Recall (Equação 2.5) e F1-Score (Equação 2.6). Precision mede a proporção de itens recomendados que são relevantes para o utilizador em relação ao total de itens recomendados, indicando quantos dos itens sugeridos realmente correspondem ao interesse do utilizador. Recall mede a capacidade do modelo recomendar todos os itens relevantes na lista de itens. F1-Score é a média harmónica entre a Precison e o Recall, permitindo encontrar um equilíbrio entre ambas. Estão são representadas pelas seguintes fórmulas:

$$Precision = \frac{VP}{VP + FP} \tag{2.4}$$

$$Recall = \frac{VP}{VP + FN} \tag{2.5}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.6}$$

onde VP corresponde ao total de verdadeiros positivos (itens relevantes recomendados corretamente), FP refere-se aos falsos positivos (itens recomendados que não são relevantes) e FN refere-se aos falsos negativos (itens relevantes que não foram recomendados).

O último tipo de métricas são as de *ranking*, que avaliam a ordem dos itens na lista resultante, como é o caso da Mean Reciprocal Rank (MRR - Equação 2.7) e Normalized Discounted Cumulative Gain (nDCG - Equação 2.8). MRR mede a qualidade do sistema através da posição onde o primeiro item relevante aparece. O nDCG compara o *ranking* obtido com o *ranking* ideal, analisando a posição de cada item recomendado com a sua posição no *ranking* ideal, permitindo perceber a qualidade da recomendação através da posição de todos os itens mais relevantes.

$$MRR = \frac{1}{U}\sum_{i=1}^{U}\frac{1}{rank_i} \tag{2.7}$$

$$nDCG = \frac{\sum_{i=1}^{p}\frac{2^{\text{rel}_i}-1}{\log_2(i+1)}}{\sum_{i=1}^{p}\frac{2^{\text{rel}_i^*}-1}{\log_2(i+1)}} \tag{2.8}$$

É importante considerar os intervalos e as situações em que cada uma das métricas de avaliação possibilita uma análise mais precisa do desempenho do sistema. As métricas de precisão, como MAE, MSE e RMSE, variam de 0 até valores infinitos, onde o desempenho do modelo é melhor quanto mais próximo estas forem de 0. O MAE é mais fácil de interpretar, uma vez que lida diretamente com erros absolutos, enquanto o MSE e o RMSE são sensíveis a erros maiores, tornando-os mais indicados quando é necessário penalizar valores de previsões muito distantes dos valores reais.

As métricas de classificação, como Precision, Recall e F1-Score, variam entre 0 e 1, com valores mais próximos de 1 a indicarem um desempenho superior do sistema. A Precision é mais útil quando o objetivo é minimizar a recomendação de itens irrelevantes, em sistemas onde a exatidão dos itens sugeridos é fundamental. O Recall é mais vantajoso em casos onde a identificação de todos os itens relevantes é crítica, como em sistemas que precisam garantir que o utilizador veja a maior quantidade possível de itens relevantes. O F1-Score oferece um equilíbrio entre essas duas métricas, permitindo avaliar o sistema quando não é possível priorizar nenhuma delas.

O MRR e o nDCG, tal como as métricas de classificação apresentadas anteriormente, variam entre 0 e 1, sendo melhores quanto mais perto de 1 estiverem. O MRR é melhor quando o foco principal está em garantir que o primeiro item relevante apareça no topo da lista, enquanto o nDCG é ideal para sistemas onde a ordem completa das recomendações tem maior impacto na experiência do utilizador.

Capítulo 3

Trabalho relacionado

Nesta secção faremos uma revisão da literatura analisada, apresentando o estado da arte e contextualizando melhor como os sistemas de recomendação têm sido aplicados tanto a plantas quanto a micróbios, explorando também o uso de grafos de conhecimento neste contexto. Primeiramente serão apresentados sistemas de recomendação criados no âmbito de auxiliar na gestão das plantas e como foram construídos. A seguir, abordamos o tópico dos grafos de conhecimentos, discutindo quais as técnicas mais atuais para a sua criação e armazenamento, assim como diferentes técnicas usadas para recomendação a partir destes. São depois revistos quais os algoritmos mais utilizados e que melhor funcionam para prospeção de interações, envolvendo micróbios. Por fim, analisamos o estado de arte de sistemas de recomendação que se aproximam ao nosso, através da utilização de micróbios, considerando várias abordagens utilizadas para a recomendação.

3.1 Sistemas de Recomendação Aplicados às Plantas

A aplicação de sistemas de recomendação na área das plantas e da agricultura tem vindo a aumentar, provando ser uma das áreas que bastante beneficia deste tipo de sistemas para a gestão e uso dos recursos eficientemente, desempenhando um papel fundamental em diversos tipos de tarefas dentro deste domínio.

Em Devan et al. (2023) [25] é proposto um modelo *ensemble* utilizando o algoritmo *Random Forest* combinado com *Logistic Regression* para recomendar fertilizantes com base na previsão de rendimentos da colheita, utilizando informações do solo, do clima e de produções anteriores.

Os autores em [20] procuraram desenvolver uma aplicação mais completa, sugerindo colheitas e fertilizantes, e classificando se as plantas se encontram doentes a partir de imagens. Foram testados vários algoritmos de aprendizagem automática, tendo o *Random Forest* apresentado melhor precisão.

Em Bandara et al. (2020) [7] foi desenvolvido um sistema de recomendação para culturas, tendo em conta fatores do solo. Os autores decidiram utilizar um modelo *ensemble* (*Naive Bayes* e *Support Vector Machine*) para fornecer a recomendação de qual tipo de colheita deve ser cultivada para determinado solo e implementaram usando *K-Means* e análise de sentimentos um sistema que permite melhorar o modelo a partir das avaliações do *feedback* dos utilizadores, identificando

a linguagem, fazendo com que o sejam ponderadas avaliações negativas consecutivas para determinada recomendação numa certa área, melhorando assim o modelo a partir de experiências reais dos utilizadores. O estudo [72] também procura aconselhar qual a colheita que se deve aplicar nos campos, considerando as necessidades do solo. Este incorporou os modelos *K-Nearest Neighbors* e *Naive Bayes* através de um modelo de *ensemble* com a técnica de Votação Maioritária.

De forma a prever plantas que podem crescer numa determinada área a partir de outras plantas, os autores em [84] criaram um sistema de reconhecimento de imagem que permite identificar a espécie através de uma das suas folhas. A previsão é feita usando filtragem colaborativa. Em Wittich et al. (2018) [97] procurou-se prever o mesmo, mas a partir de características do território.

Com o objetivo de recomendar um meio de controlo contra pragas, alguns sistemas foram desenvolvidos. Os autores em [85] criaram um sistema capaz de reconhecer doenças através de imagens, recomendando inseticidas capazes de as neutralizar. Em Pudumalar et al. (2018) [71] recomendaram pesticidas químicos utilizando *Decision Trees* e *Naive Bayes* como algoritmos de classificação.

Os estudos demonstram uma elevada tendência para o uso de modelos *ensemble*, combinando diferentes algoritmos, dos quais se destaca o uso de *Naive Bayes*. Além disso, técnicas avançadas como o reconhecimento de imagem e a análise de sentimentos foram integradas a esses modelos, aumentando a capacidade dos sistemas de recomendação de se adaptarem a diferentes condições e alterações no meio, usando dados reais para otimizar as previsões.

Recentemente, foi lançada uma aplicação web, o BioProtectionPortal [1], que oferece produtos biopesticidas à base de micro e macro-organismos, extratos botânicos e semioquímicos, destinados ao controle de pragas em colheitas específicas. A ferramenta considera os produtos registados em cada país, fornecendo opções adequadas para o combate a determinadas pragas agrícolas.

Contudo, até onde sabemos, não foram encontrados estudos que recomendassem novos micróbios a patógenos de plantas.

3.2 Grafos de Conhecimento em Sistemas de Recomendação

Na era atual a incorporação de grafos do conhecimento em áreas biológicas tem-se tornado cada mais evidente [57, 83, 52, 88], demonstrando ser uma das abordagens mais eficazes para guardar o conhecimento, conservando as ligações entre entidades. Esta prática revela-se eficiente para a aplicação do conhecimento em diversas tarefas de aprendizagem automática [32]. Este conhecimento pode ser recolhido de vários tipos de dados, estruturados, semiestruturados ou não estruturados. Quando os dados são estruturados facilmente se identificam entidades e relações. Por outro lado, quando são semiestruturados ou não estruturados é necessário aplicar técnicas complexas de prospeção de interações que permitam concluir essas relações [82].

O desenvolvimento de um grafo de conhecimento tornou-se então um dos principais desafios, os avanços nas técnicas de processamento de linguagem natural, fazem com que o estado de arte

[1] https://bioprotectionportal.com/

se encontre numa evolução constante. Por isso o autor em [82] delineou, a partir da análise de outros estudos, uma sequência de procedimentos generalizada para o desenvolvimento de um grafo de conhecimento, dividida em 6 passos, com possíveis subtarefas: (i) Identificação dos dados, qual a fonte dos dados e quais os tipos de dados retirados; (ii) Construção da ontologia, no caso da existência de uma; (iii) Extração do conhecimento, identificar e retirar as entidades relevantes nos dados, deduzir as relações existentes entre essas entidades e retirar atributos que caracterizam essas entidades; (iv) Processamento do conhecimento extraído, analisar o conhecimento extraído, removendo duplicados e agrupando relações semelhantes; (v) Construção do grafo de conhecimento, guardar este no modelo de dados mais adequado à sua visualização e disponibilidade para uso; (vi) Manutenção, avaliar o grafo de conhecimento e constante atualização dos dados.

A alocação do grafo de conhecimento numa base de dados de grafos própria é importante para garantir uma melhor administração e visualização. O estudo [63] comparou as quatro melhores bases de dados de grafos classificadas em 2022, concluindo que a base de dados Neo4j em termos de eficiência, carregamento de nós e tempo de consulta, proporciona um desempenho superior.

O uso de grafos de conhecimento em sistemas de recomendação tem proporcionado uma mais-valia na eficácia de recomendação de itens. Uma das abordagens mais comuns e aplicadas nesta área é a utilização de *embeddings* do grafo de conhecimento.

Os embeddings do grafo de conhecimento representam as entidades e relações do grafo em um espaço vetorial de baixa dimensionalidade. Através dessa representação é possível capturar informações sobre as relações de forma mais eficaz e clara. Segundo Zhang et al. (2024) [104] estes podem ser classificados em 2 abordagens diferentes, distância translacional e correspondência semântica. A tabela 3.1 oferece uma visão geral das diferentes categorias e de métodos existentes para cada uma delas.

Tabela 3.1: Visão geral de métodos de embeddings

Categoria	Método	Função de Pontuação	Artigo
Distância Translacional	TransE	$\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_{1/2}$	[13]
	TransH	$\| \left(\mathbf{h}-\mathbf{w}^T\mathbf{r}\mathbf{h}_\mathbf{w}\right)+\mathbf{r}-\left(\mathbf{t}-\mathbf{w}^T\mathbf{r}\mathbf{t}_\mathbf{w}\right) \|_2^2$	[94]
	RotatE	$\|\mathbf{h}\circ\mathbf{r}-\mathbf{t}\|_2^2$	[81]
Correspondência Semântica	DistMult	$h^T\text{diag}(\mathbf{r})t$	[100]
	ComplEx	$\text{Re}(h^T\text{diag}(\mathbf{r})t)$	[86]
	SimplE	$\frac{1}{2}\left(\langle\mathbf{h}_h,\mathbf{r},\mathbf{t}_t\rangle+\langle\mathbf{h}_t,\mathbf{r}^{-1},\mathbf{t}_h\rangle\right)$	[45]

Os métodos de embeddings de distância translacional utilizam funções de pontuação baseadas na distância, tratando as relações entre entidades como translações num espaço vetorial, para calcular a semelhança entre entidades [48]. Exemplo de métodos para esta categoria são TransE, TransH e RotatE. O TransE é o mais simples interpretando as relações como translações lineares entre entidades, representando tanto as relações como entidades no mesmo espaço. O TransH atribui a cada relação um hiperplano específico, projetando ambas as entidades (cabeça e cauda) nesse hiperplano. O RotatE baseia-se no TransE, mas utiliza rotações em vez de translações, para

definir cada relação.

Os métodos de embeddings de correspondência semântica utilizam funções de pontuação baseadas em similaridade para avaliar a relação entre entidades, modelando as relações como interações entre as características latentes das entidades [48]. Exemplos de métodos são o DistMult, ComplEx e SimplE. O DistMult representa as relações entre entidades como matrizes diagonais, onde os elementos da diagonal correspondem às características latentes das entidades. A similaridade entre um par de entidades é calculada através do produto escalar das suas representações vetoriais, multiplicado pelo valor na matriz diagonal da relação. O ComplEx estende o DistMult ao representar entidades e relações como vetores complexos, permitindo a captura de relações simétricas e assimétricas. O SimplE utiliza duas representações para cada entidade, dependendo se é cabeça ou cauda num triplo, calculando a similaridade combinando a interação entre *cabeça, relação, cauda* com a interação entre *cauda, relação inversa, cabeça*.

Vários estudos têm explorado o uso de embeddings de grafos de conhecimento como base para fornecer recomendações. Em Geleta et al. (2021)[32] os autores reuniram dados de diferentes conjuntos de dados, criando um grafo de conhecimento com informações, compostos, doenças e genes, pertinentes ao desenvolvimento de novas substâncias. Utilizaram uma técnica chamada aprendizagem representacional para representar cada nó no grafo por meio de mapeamento para um ponto num espaço vetorial de baixa dimensão, preservando assim a topologia do grafo. Em Yang et al. (2018)[101] é implementado um modelo de recomendação baseado em *GAN*. Este aprende as representações iniciais de filmes e dos utilizadores com base em *embeddings* de conhecimento e de *tags* utilizando o modelo Metapath2Vec e o modelo Word2Vec. Em seguida, treina um gerador, que procura filmes relevantes, e um discriminador, que procura filmes irrelevantes. O artigo conclui que o uso de *embeddings* de grafos de conhecimento pré-treinados melhora o desempenho do modelo. Os autores em [18] propuseram o modelo KHGCN, focado em sistemas de recomendação personalizados, capaz de aprender relações entre entidades de um grafo, ao mesmo tempo que filtra dados irrelevantes. O modelo extrai os *node embeddings* no grafo, aprendendo a sua estrutura hierárquica através de várias camadas de cápsulas, usando *Transformer GATs*, partes de roteamento residual e um mecanismo de atenção para robustecer a agregação do grafo de conhecimento.

Existem outras abordagens, como a baseada em Conexões, que utiliza padrões nas ligações do grafo de conhecimento para guiar as recomendações [105], e como a baseada em Propagação, que se concentra na propagação e difusão de informações pelo grafo para capturar a semântica inerente e as ligações entre entidades, tirando partido da estrutura e das relações, mas que, no entanto, têm dificuldades quando existem relações complexas ou ruído [93].

3.3 Prospeção de Interações

Atualmente ainda não existe uma base de dados que contenha relações entre micróbios e patógenos de plantas, residindo a sua informação em artigos científicos, o que torna o seu acesso bastante complicado e demorado. Na secção 3.2 referenciou-se que a procura dessas entidades e extração

das suas relações em dados não estruturados necessita o uso de ferramentas complexas, como é o caso de reconhecimento de entidades nomeadas e extração de relação. O estado de arte destas ferramentas está atualmente dominado por modelos híbridos de aprendizagem profunda e linguagem pré-treinada [64].

A aplicação do reconhecimento de entidades nomeadas e extração de relação geralmente é um processo contínuo e interligado [64], onde o reconhecimento de entidades nomeadas identifica e classifica as diferentes entidades, enquanto a extração de relação analisa as interações/relações entre essas mesmas entidades, no mesmo texto.

Ambas as tarefas são bastante usadas na área da microbiologia, onde a informação é maioritariamente guardada sobre a forma de literatura. Os autores em [98] criaram uma base de dados que contivesse as relações entre micróbios e doenças humanas, extraindo o conhecimento de literatura. Na análise ao estado de arte, estes concluíram não existir nenhuma ferramenta de reconhecimento de entidades nomeadas para micróbios. Utilizaram o LINNAEUS, um sistema de identificação de nomes para literatura biomédica baseada num dicionário, para encontrar nomes de micróbios e usaram DNorm para reconhecer e normalizar doenças no texto. Após a identificação aplicaram duas ferramentas distintas para extração de relação, PKDE4J, modelo baseado em regras, e BERE, um modelo pré-treinado, originalmente construído para reconhecer interações substância-substância, que utiliza aprendizagem de árvores latentes e técnicas de autoatenção. De forma a avaliarem o desempenho, criaram um conjunto de dados manualmente curado, para servir de base de comparação. Concluíram que com o modelo BERE obtinham melhores resultados com um F-Score de 0.73. Um problema que encontraram foi a falta de capacidade por parte do reconhecimento de entidades nomeadas de reconhecer abreviaturas de micróbios existentes na literatura.

Em Karkera et al. (2023)[44] foi proposto encontrar relações entre micróbios e doenças extraindo informação da literatura biomédica, a partir de processamento de linguagem natural usando modelos de linguagem pré-treinados. Usaram 3 modelos generativos diferentes, GPT-3, BioMedLM e BioGPT e 5 modelos discriminativos BERT, PubMedBERT, BioMegatron, BioLINKBERT e BioClinicalBERT. O objetivo era perceber quais os modelos que melhor se comportavam na classificação de relações, tendo em conta a aprendizagem por que estes tivessem passado anteriormente. Para isso foram testados os modelos generativos sem e com aprendizagem, e os modelos discriminativos afinados. Os modelos generativos sem qualquer ajuste demonstraram resultados bastante pobres, necessitando de modelos de domínio específico, destacando-se o GPT-3 quando afinado. Por outro lado, os modelos discriminativos demonstraram obter melhor desempenho, atingindo F-Score, Precision e Recall acima de 0.8.

Os modelos BiLSTM-CRF, BioBERT-CRF e PubMedBERT-CRF de reconhecimento de entidades, bem como o BERT-GT e o PubMedBERT de extração de relação, foram usados para avaliar um conjunto de dados biomédicos em [56].

O uso destes tipos de métodos no reconhecimento de entidades nomeadas nem sempre é a melhor forma de extrair as entidades pretendidas, o uso de *tags* pode por vezes não permitir selecionar entidades específicas, ou agrupar estas em categorias menos específicas [70]. Para isso

existem os métodos baseados em dicionários que usam taxonomias próprias a partir de uma lista de entidades. Um exemplo do seu uso é no artigo [68], onde este foi empregue para não haver más anotações entre doenças e micróbios, utilizando uma lista específica de bactérias e vírus.

3.4 Sistemas de Recomendação de Micróbios

A tabela 3.2 mostra diversos trabalhos desenvolvidos na recomendação aplicada a micróbios ou patógenos. Esta contém informação sobre o ano do estudo, o que foi considerado como utilizador e como item, o tipo de recomendação utilizada, e o conjunto de dados utilizado, sabendo se este está ou não disponível para utilização.

Tabela 3.2: Estudos anteriores de previsões a partir de micróbios ou fatores de plantas

Artigo	**Ano**	**Utilizador**	**Item**	**Filtragem**	**Conjunto de Dados**	**Disponível**
Li et al. [50]	2014	Patógenos	Proteínas Hospedeiras	Colaborativa	PHISTO	Não
Huang et al. [36]	2017	Doenças Humanas	Micróbios	Colaborativa	HMDAD	Sim
Long et al. [54]	2020	Micróbios	Medicamentos	Híbrida	MDAD, aBiofilm & DrugVirus	Não
Liu et al. [53]	2021	Doenças Humanas	miRNAs	Colaborativa	HMDDv2.0	Sim
Xu et al. [99]	2022	Doenças Humanas	Micróbios	Colaborativa	HMDAD	Sim
Medvedeva et al. [60]	2023	Bactérias	Péptidos antimicrobianos	Colaborativa	DBAASP	Não

Através da análise à tabela 3.2 podemos ver que os sistemas de recomendação envolvendo patógenos e micróbios, baseiam-se no sistema tradicional de utilizador e item. Em alguns estudos [36, 54, 99] são recomendados micróbios a doenças humanas. Outros recomendam efeitos de proteínas hospedeiras a patógenos [50], miRNAs a doenças humanas [53] e efeitos de determinados péptidos antimicrobianos a micróbios [60].

A filtragem colaborativa é a mais usada em todos os estudos. Devido à maioria dos conjuntos de dados não conterem características sobre os itens, uma abordagem utilizando filtragem baseada em conteúdo seria bastante mais desafiadora de ser implementada efetivamente. No entanto, podia trazer vantagens, como a recomendação de itens quando não se tem informações de *ratings* para esses itens.

A integração da fatorização de matrizes surge como um método recorrente e poderoso em vários artigos [50, 53], mostrando a sua versatilidade em diversas aplicações. Já em Xu et al. (2022)[99] optaram por utilizar fatorização matricial não negativa. A eficácia deste método no

tratamento de dados biológicos complexos sublinha a sua importância na descoberta de padrões significativos. Os métodos baseados na vizinhança foram usados como principal método em [36, 60]. A utilização de um método baseado em grafos para melhorar a qualidade das previsões, foi testada também em [36]. O artigo [54] utiliza uma abordagem distinta empregando uma Rede Neural Convolucional em Grafos para lidar com a complexidade das interações entre micróbios e medicamentos. Esta permite integrar diversas fontes de informação biológica, incluindo redes de similaridade entre micróbios e medicamentos, bem como interações micróbio-medicamento, ao incorporar uma camada de Campo Aleatório Condicional e um mecanismo de atenção.

Neste trabalho pretendemos utilizar sistemas de recomendação de grafos, para recomendar micróbios a patógenos das plantas. Como podemos observar na tabela 3.2, não existem dados disponíveis para testar sistemas de recomendação de micróbios para patógenos das plantas, como tal, teremos de criar o nosso próprio conjunto de dados.

Capítulo 4

Metodologia

O objetivo deste trabalho é a criação de uma conjunto de dados de *feedback* implícito para a recomendação de agentes de controlo biólogico, mais especificamente micróbios, para doenças de plantas e a sua integração em um grafo de conhecimento de forma a ser usado num sistema de recomendação. Este conjunto de dados pretende seguir a estrutura padrão dos conjuntos de dados utilizados em sistemas de recomendação de <utilizador,item,*rating*>, no entanto os utilizadores são os patógenos de plantas, os itens os micróbios e os *ratings* o número de artigos que cada micróbio é referenciado com um patógeno. Para a sua criação é seguida a metodologia LIBRETTI proposta em [10] para criação de conjuntos de dados de *feedback* implícito a partir de literatura científica. O trabalho está dividido em quatros etapas distintas: Recolha de dados; Criação do conjunto de dados; Construção do grafo de conhecimento; Sistema de recomendação. A Figura 4.1 representa uma visão geral da metodologia aplicada, detalhando os processos usados em cada uma das etapas.

4.1 Fontes de Dados

A inexistência de uma base de dados específica para patógenos de plantas e micróbios constitui um dos principais desafios deste projeto, obrigando à recolha de dados a partir de múltiplas fontes com o intuito de criar um conjunto de dados o mais abrangente e completo possível.

Os nomes dos patógenos são obtidos através da PHI-base, uma base de dados composta por informações extraídas de artigos de investigação validados, que documentam genes influenciando as interações entre patógeno e hospedeiro [87]. Esta base de dados abrange diferentes tipos de hospedeiros, incluindo plantas, animais e humanos, e permite compreender como a modificação ou remoção experimental de genes específicos de patógenos pode reduzir a severidade ou efetividade do patógeno causar a doença. A PHI-base contém aproximadamente 28 840 registos, com informações detalhadas sobre o tipo de gene (nome, sequência e função), sobre os patógenos (nome da espécie e estirpe), sobre os hospedeiros (tipo e espécie), bem como o nome da doença associada.

A obtenção de dados relativos a micróbios foca-se em duas categorias diferentes, procariontes e fungos. Para cada uma dessas categorias, são utilizados dois conjuntos de dados distintos: um

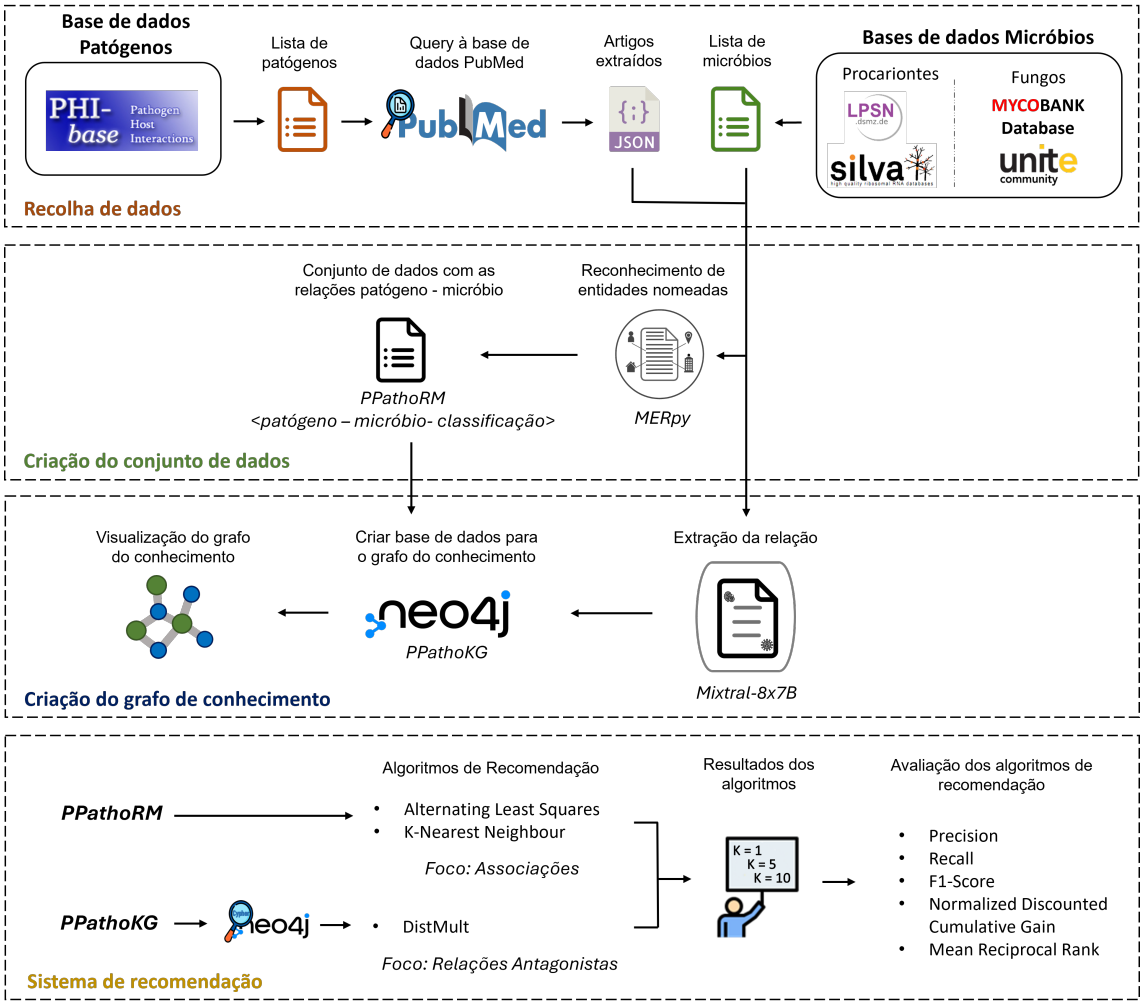

Figura 4.1: Visão geral da metodologia

contendo os nomes oficiais de cada micróbio e outro com os nomes mais comuns utilizados nos estudos científicos.

No caso dos procariontes, os nomes oficiais foram extraídos da base de dados LSPN, disponibilizada em formato CSV, contendo uma lista de 29 765 grupos taxonómicos distintos [69]. Para os nomes mais comuns usados em estudos, recorreu-se à base de dados SILVA, utilizando a versão v138.1, disponibilizada em formato FASTA, especializada em sequências de RNA ribossómico de microrganismos [73, 102, 59].

Para os fungos, as informações sobre os grupos taxonómicos foram obtidos a partir da base de dados MycoBank, que contém 594 793 grupos taxonómicos e respetivas associações [75, 24, 74]. Os nomes utilizados nos estudos foram extraídos da base de dados UNITE, que inclui todos os espaçadores internos transcritos de cada micróbio, bem como os seus respetivos nomes [1].

Extraímos da base de dados da NCBI [65] um conjunto de dados taxonómicos, contendo diversos grupos taxonómicos para micróbios, incluindo nomes científicos e sinónimos, agrupados por um id. Foi utilizado o pacote taxonomizr [76] da linguagem R para aceder e extrair os dados, sendo esta ferramenta concebida propositadamente para trabalhar com a estrutura da base de dados da NCBI. A sua capacidade de acesso local à base de dados oferece maior eficiência no processamento de grandes volumes de dados, como é o caso da extração de múltiplos grupos taxonómicos, o que a torna uma escolha comum em vários estudos recentes [8, 23].

4.2 Recolha de Dados

Uma das características da recolha de dados de diversas fontes consiste na variedade e heterogeneidade destes dados, tanto a nível descritivo como estrutural. Por exemplo, estes dados apresentam formatos variados e utilizam convenções de nomeação distintas, o que pode resultar em diversos desafios de integração, como a duplicação de dados, inconsistência nos dados e a escalabilidade.

Primeiro, foram tratados os dados relativos aos patógenos, de forma a processá-los para resultarem numa lista de patógenos. Os dados contêm uma enorme variedade de espécies, com cerca de 288 patógenos, dependendo do tipo de hospedeiro, filtrando-se por aqueles que são referentes a tipos de plantas, nomeadamente, eudicotiledôneas, monocotiledóneas, plantas com flores e plantas com sementes. As restantes características são descartadas por não apresentarem qualquer relevância para o estudo, eliminando simultaneamente as espécies duplicadas.

Quanto aos dados dos micróbios, cada conjunto de dados teve um tratamento diferente. Os ficheiros CSV, referentes aos grupos taxonómicos oficiais, foram tratados primeiro por conterem mais

informações sobre cada um dos micróbios, assim como um maior volume de informação. Os dados relativos aos procariontes, extraídos da base de dados LSPN, continham vários níveis taxonómicos, como género, espécie e subespécie, apresentando tanto nomes oficiais como sinónimos. Apenas os grupos taxonómicos que abrangiam tanto o género como a espécie foram mantidas, descartando as restantes. Os sinónimos foram tratados como grupos taxonómicos distintos. Os dados dos fungos, obtidos da base de dados MycoBank, continham apenas dois níveis taxonómicos, género e

espécie, mantendo-se os grupos taxonómicos que continham ambos. Os ficheiros FASTA, contendo sequências únicas de cada espécie, foram processados utilizando a ferramenta Biopython [17]. Para ambos os conjuntos de dados, SILVA e UNITE, o procedimento foi o mesmo, procedendo-se à extração dos vários grupos taxonómicos presentes e removendo os duplicados. Após a extração dos nomes reparou-se na existência de alguns destes que não deveriam constar na lista por não acrescentarem valor, generalizando e não se referindo a uma espécie em concreto, filtrando todos os que a espécie se denominasse por *bacterium, metagenome, symbiont* e *endosymbionts*.

Para cada um dos grupos taxonómicos presentes nas listas, foram procurados os seus sinónimos no conjunto de dados da NCBI e estes foram adicionados a cada uma das listas. No final desta etapa temos uma lista de nomes e sinónimos de patógenos e micróbios que nos permite passar à fase seguinte, que consiste na recolha de artigos científicos relacionados com estes patógenos.

De seguida foi usada a API da Entrez Programming Utilities (E-utilities) [29] para extrair artigos científicos do PubMed. Para a *query* foram utilizados os nomes dos patógenos constantes na lista gerada no passo anterior, extraindo o máximo de artigos que existissem na base de dados sobre cada um destes patógenos, com um limite máximo de 10 000 imposto pela ferramenta utilizada, com data de publicação entre 2003 e 2024. A utilização de aspas (") antes e depois do nome do patógeno garante que a procura seja mais precisa, não divagando para outros patógenos com o mesmo nome de espécie. Foi guardado o PubMedID (identificador do PubMed), o título, o resumo, os autores, o ano de publicação e o DOI $Digitalobjectidentifier$ de cada artigo, conjuntamente num ficheiro JSON.

Foi então criada uma base de dados em SQLite, com as informações recolhidas, permitindo associar os patógenos a cada artigo e prevenir a duplicação de artigos na procura destes. A base de dados é constituída pela tabela dos patógenos, a tabela dos artigos (guardando apenas o PubMedID, os autores, o ano e o DOI) e a tabela associativa patógenos-artigos, uma vez que um artigo pode abordar mais do que um patógeno e cada patógeno pode ser mencionado em vários artigos. Patógenos existentes na lista, mas sem artigos associados não foram incluídos na base de dados. A estrutura da base de dados pode ser consultada na Figura 4.2

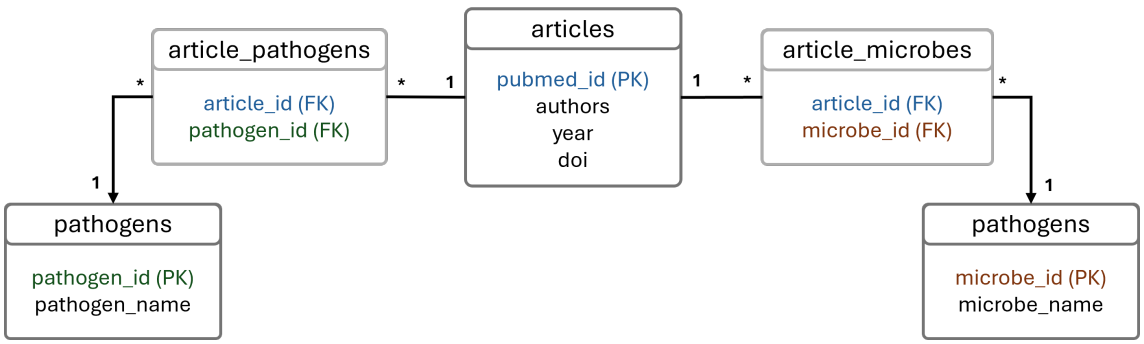

Figura 4.2: Estrutura da base de dados

4.3 Criação do Conjunto de Dados Patógeno-Micróbio-*Rating*

Construindo sobre os dados recolhidos, a segunda etapa centra-se na criação de um conjunto de dados robusto, com todas as associações patógenos-micróbios. Na fase anterior recolheram-se artigos científicos relacionas com os patógenos de plantas. Nesta fase, o objetivo é utilizar a lista de micróbios produzida anteriormente, para extrair a partir do texto desses artigos, os micróbios que estão associados aos patógenos.

Técnicas de reconhecimento de entidades nomeadas permitem extrair entidades de um texto, fornecendo a sua localização exata e o tipo de entidade. A maioria dos modelos são treinados de forma a interpretar o texto e identificar múltiplos tipos de entidades. Devido à especificidade das entidades que se pretendem identificar, não existe nenhuma ferramenta treinada pronta para identificar unicamente micróbios. Posto isto, optou-se for utilizar uma ferramenta mais simples, a MERpy [22] que funciona a partir de um dicionário de entidades, utilizando a lista de micróbios como léxico. Para determinar a qualidade das anotações, foram selecionados aleatoriamente 70 artigos, separados por 6 anotadores diferentes, onde cada um analisou 15 artigos, tendo 4 em comum com outro anotador.

A pesquisa de micróbios é realizada percorrendo todos os artigos obtidos, abrangendo as suas secções, título e resumo. Para cada artigo, é criado um ficheiro JSON, utilizando o PubMedID como identificador. Este ficheiro contém os micróbios identificados, o número de vezes que cada um é mencionado no texto, bem como a sua localização exata, discriminada por secção. Os micróbios encontrados são guardados na base de dados e associados ao artigo em que foram encontrados.

Finalmente, foi usada a base de dados já completa com todos os patógenos e micróbios existentes em cada artigo, para se proceder à contagem de artigos, em que cada patógeno era mencionado em conjunto com cada micróbio, para criar o conjunto de dados PPathoRM na forma de <utilizador,item,*rating*>. Os utilizadores são os patógenos, os itens os micróbios e o *rating* o número de artigos que um patógeno foi mencionado com um micróbio. Usámos o conjunto de dados da NCBI para agrupar os sinónimos de cada entidade sob o seu nome cientifico, garantindo a normalização dos grupos taxonómicos adicionados anteriormente para melhorar a procura de entidades no texto.

4.4 Construção do Grafo de Conhecimento

O conjunto de dados destaca apenas a existência de uma associação entre patógeno e micróbio, considerando a frequência com que estes são referenciados conjuntamente. No entanto, o tipo de associação pode fornecer informação mais explícita de como estes dois se relacionam, permitindo-nos tirar efetivamente conhecimento sobre o fator de inibição de um micróbio sobre um patógeno. Assumir que a simples coocorrência de duas entidades num artigo implica uma relação inibitória por parte do micróbio sobre o patógeno é inadequado. A referência pode considerar micróbios bastante promissores identificados a partir de outros estudos, mas que, após testes, se revelaram incapazes de inibir os patógenos específicos ou, em alguns casos, até apresentaram efeitos

contrários. Por conseguinte, é essencial considerar o tipo de associação descrita em cada artigo, uma vez que um micróbio pode não só falhar em inibir o patógeno, como também pode cooperar com este. Isto permite assim criar verdadeiramente um grafo de conhecimento, que reflete diferentes relações que possam existir entre patógeno e micróbio.

Para identificar o tipo de relação existente entre cada par patógeno-micróbio, procedeu-se à extração de relações. Foi utilizado um Modelo de Linguagem de Grande Escala, o Mixtral-8x7B-Instruct-v0.1 [41], capaz de interpretar textos e responder a questões, através da API do Hugging Face [39]. Sendo um modelo onde o *output* é de resposta aberta, podendo divergir entre os tipos de relações encontradas, definimos 4 *tags* como respostas possíveis:

- **Antagonist**: O micróbio tem capacidades inibitórias para determinado patógeno.
- **Agonist**: O micróbio coopera com o patógeno, ajudando a ativá-lo.
- **NoRelation**: Não existe qualquer tipo de relação entre micróbio e patógeno.
- **Unknown**: Não foi possível verificar a existência de uma relação, o tipo de relação encontrada não se enquadra em nenhuma das *tags* anteriores ou o *output* do modelo não resultou em apenas uma palavra.

Como *input* para o modelo, foi utilizado um texto criado por nós, que incluía o objetivo, a tarefa, o tipo de *output* pretendido, um exemplo, o resumo e o patógeno e o micróbio presente, cuja relação se pretendia identificar. O *input* utilizado pode ser consultado na Figura 4.3.

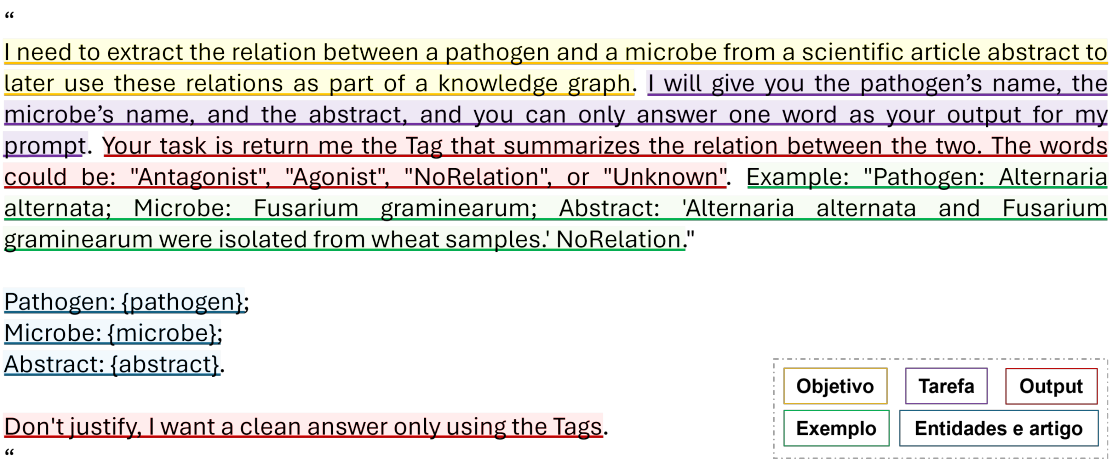

Figura 4.3: Input usado no modelo de Extração de Relação

As relações foram agrupadas por *tag* e por par patógeno-micróbio. A variedade de artigos para cada patógeno, especialmente em função do ano de publicação, apresentam uma possibilidade de que existam diferentes relações entre os mesmos pares patógeno-micróbio dependendo do artigo. A capacidade dos testes e dos estudos realizados com o passar do tempo evoluí, levando a novas descobertas não possíveis antes. No caso de haverem mais do que uma relação consideramos a

relação presente no artigo mais recente. Além disso, incluímos um *rating* adicional relativo ao número de artigos que apresentam essa relação mais atual.

Com as relações extraídas entre patógenos e micróbios, foi possível criar o grafo de conhecimento, PPathoKG. Para isso utilizou-se uma base de dados em Neo4j, inserindo cada patógeno e micróbio como um nó, sendo o seu nome uma propriedade e o tipo de entidade a sua categoria, e cada relação como uma aresta, possuindo 2 propriedades: *Rating*, todos os artigos encontrados entre aquelas duas entidades; *Rating* da relação, todos os artigos que continham aquela relação entre as duas entidades. Foi realizada a normalização dos grupos taxonómicos das entidades tal como feito na Secção 4.3.

4.5 Sistema de Recomendação

A partir das etapas anteriores obtivemos o conjunto de dados PPathoRM, através do reconhecimento de entidades nomeadas, e o grafo de conhecimento PPathoKG, usando a informação da extração da relação com o conjunto de dados.

A implementação de um algoritmo de recomendação tem dois objetivos diferentes, determinar a qualidade do conjunto de dados para aplicação em sistemas de recomendação, e se a utilização do grafo de conhecimento, com a adição das diferentes relações entre patógeno e micróbio, é capaz de melhorar as recomendações de micróbios com capacidades inibitórias para patógenos.

Ambos foram avaliados utilizando métodos de avaliação *offline* considerando a qualidade da lista ordenada de itens recomendados. Do conjunto de métricas existentes para fazer este tipo de avaliação foram selecionadas a Precison@k (Equação 2.4), Recall@K (Equação 2.5), F1-Score@K (Equação 2.6), Mean Reciprocal Rank@K (Equação 2.7) e nDCG@K (Equação 2.8), sendo K o número de itens recomendados na lista.

4.5.1 Qualidade do Conjunto de Dados

Para avaliar as recomendações no conjunto de dados, utilizámos a biblioteca *Fast Python Collaborative Filtering for Implicit Datasets* [30]. Esta biblioteca foi concebida para fornecer uma variedade de algoritmos de recomendação amplamente utilizados em filtragem colaborativa, especificamente para conjuntos de dados baseados em *feedback* implícito. Aplicámos um algoritmo baseado em memória, o *K-Nearest Neighbors* (KNN) usando a Similaridade do Cosseno entre itens, e um algoritmo baseado em modelos, o *Alternating Least Squares* (ALS). O KNN representa os itens num vetor e usa Similaridade do Cosseno para calcular os K itens mais próximos aos itens que o utilizador já interagiu. O ALS é uma técnica de fatorização de matrizes que otimiza iterativamente os fatores latentes de utilizadores e itens para minimizar o erro de previsão.

O conjunto de dados foi dividido utilizando o *cross-validation* com 5 separações, em que a divisão é de 80% para o conjunto de treino e 20% para o conjunto de teste. Como parâmetros para o KNN, o número de vizinhos considerado foi 5 (K=5). Para o ALS, foram utilizados 150 fatores latentes (*factors*=150).

Criamos uma variação do PPathoRM, o PPathoRM20 incluindo apenas os patógenos que continham mais de 20 *ratings*. Este conjunto de dados foi criado parar melhorar a precisão e a robustez das recomendações, ao garantir que os dados são mais representativos, prevenindo o *cold start* para utilizadores.

4.5.2 Recomendações Baseadas em Relações

A procura de micróbios inibitórios para determinados patógenos foca-se na descoberta daqueles que tenham relações antagonistas com estes. O uso do grafo de conhecimento, contendo as relações entre as diferentes entidades, procura discriminar as diferentes relações e permitir que as recomendações sejam apenas aquelas com propriedades inibitórias.

A extração dos dados da base de dados no Neo4j foi feita utilizando a linguagem de procura desta, Cypher, extraindo os vértices, contendo id da entidade de origem, a relação e o id da entidade de destino e os nós contendo o id da entidade, o tipo de entidade (patógeno ou micróbio) e o nome da entidade.

O método aplicado está representado em Figura 4.4. Usamos o modelo de recomendação DistMult [100], da biblioteca *Fast Graph Representation Learning with PyTorch Geometric* [28], com 64 canais ocultos. O DistMult é um modelo de fatorização de tensores, em que cada entidade e relação são representadas como um vetor no espaço de embeddings. Utilizamos depois os embeddings das entidades e das relações, focando no patógeno em questão e na relação antagonista para criar um vetor de pesquisa. Esse vetor é depois aplicado nos embeddings das entidades através do produto escalar para calcular a pontuação de cada triplo (p,r,m). No processo de otimização foi utilizado o método com o otimizador Adam, com uma taxa de aprendizagem de 0.001, e treinado o modelo com 20 *epochs*. O modelo utiliza como *input* os vértices do grafo, com a origem, relação e destino, e o número de nós.

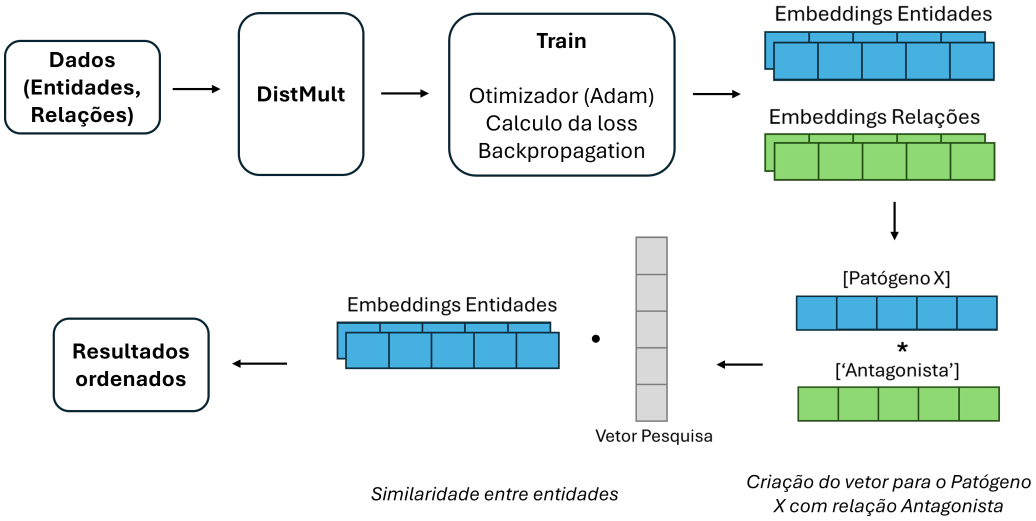

Figura 4.4: Método utilizado para aplicação dos embeddings produzidos pelo DistMult para recomendação de micróbios.

Para termos uma base de comparação e determinar se a integração das relações beneficia as recomendações de micróbios com capacidades inibitórias de determinados patógenos, criámos um sistema de recomendação para servir como base de comparação utilizando o PPathoRM e o algoritmo ALS, mantendo-se os parâmetros da etapa anterior. Nesta abordagem, os itens relevantes considerados são apenas as associações entre entidades que representem relações de antagonismo, contrastando com o sistema de recomendação anterior, onde foram considerados relevantes quaisquer tipos de associações.

Foram excluídos todos os patógenos que apresentavam menos de duas relações antagonistas, garantindo que o modelo tivesse pelo menos uma dessas associações disponíveis para o treino e uma para ser avaliada no teste. Além disso, assegurámos que o conjunto de treino e teste fosse o mesmo para ambos os algoritmos, de forma a permitir uma comparação direta entre eles. Utilizámos uma divisão simples dos dados, com 80% para treino e 20% para teste, considerando como itens relevantes exclusivamente as relações antagonistas.

Capítulo 5

Resultados e Discussão

Este capítulo está dividido em 4 partes. Na primeira, é apresentada uma análise das listas de patógenos e micróbios obtidas, bem como dos artigos extraídos do PubMed. A segunda parte apresenta os resultados da avaliação da ferramenta de reconhecimento de entidades nomeadas na identificação de micróbios presentes nos artigos científicos e o conjunto de dados criado a partir desta ferramenta. A terceira parte apresenta as relações extraídas através da extração de relação e uma análise ao grafo de conhecimento. Por fim, a quarta parte apresenta os resultados da validação do conjunto de dados, utilizando os algoritmos ALS e KNN, e descreve os resultados do sistema de recomendação desenvolvido com base no grafo de conhecimento, comparando-o com o sistema de recomendação que utiliza apenas o conjunto de dados e o algoritmo ALS, no contexto da recomendação de relações antagonistas.

5.1 Recolha de Dados

5.1.1 Listas de Patógenos e Micróbios

Obtivemos no total 462 803 grupos taxonómicos de micróbios (considerando sinónimos), onde 432 823 são fungos e 29 979 são procariontes. Para os patógenos foram obtidos 389 grupos taxonómicos diferentes, referentes a 161 entidades únicas. A tabela 5.1 e 5.2 permitem visualizar uma amostra da lista de micróbios e de patógenos, respetivamente.

Tabela 5.1: Amostra da lista de micróbios

Grupo Taxonómico	Categoria
Amylostereum areolatum	Fungi
Allomuricauda algicola	Prokaryotes
Alteribacter lacisalsi	Prokaryotes

Tabela 5.2: Amostra da lista de patógenos

Grupo Taxonómico
Agrobacterium tumefaciens
Alternaria alternata
Bipolaris maydis

5.1.2 Artigos Científicos

A lista de patógenos permitiu-nos a extração de artigos de literatura científica da base de dados do PubMed. Foram no total obtidos 104 967 artigos diferentes, sendo alguns destes comuns a mais do que um patógeno. A figura 5.1 apresenta o número de artigos por ano. É possível observar

que existe um crescente aumento do número de artigos ao longo dos anos, atingindo um máximo em 2021 com 6 905 artigos referentes aos patógenos na nossa lista. A extração dos artigos foi realizada no mês de Maio de 2024, justificando o baixo número de artigos para 2024. Estes dados evidenciam um crescimento nas publicações científicas sobre patógenos, indicando um crescente interesse e uma maior quantidade de estudos nesta área.

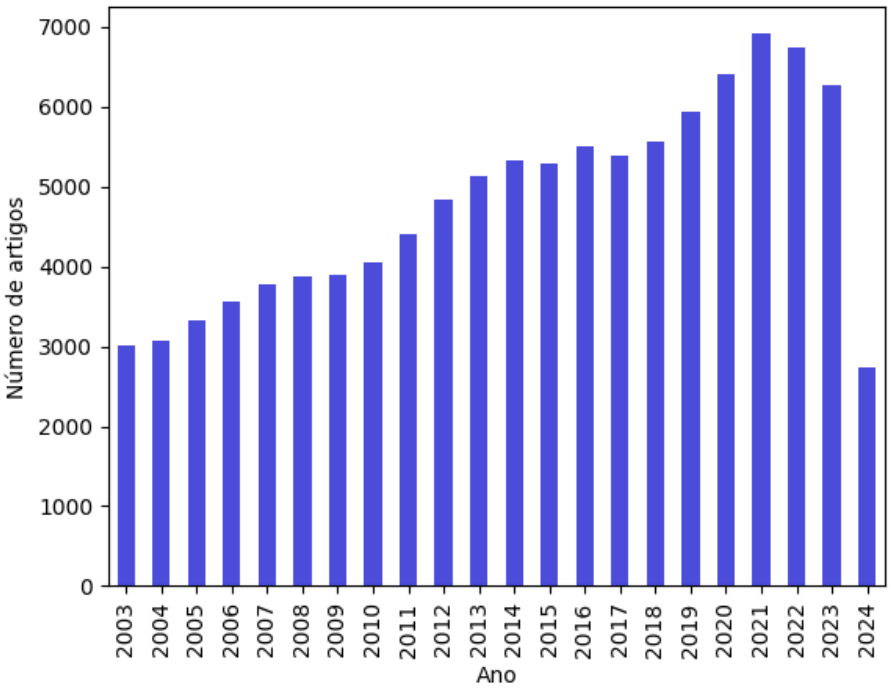

Figura 5.1: Distribuição de artigos científicos por ano.

A figura 5.2 apresenta a distribuição do número de artigos por patógeno, ordenada pelo número de artigos e agrupando os sinónimos sob os respetivos nomes científicos. Observa-se que existe um número baixo de patógenos que contêm muitos artigos científicos, sendo a média de 471.99 artigos por patógeno. Isto demonstra a existência de patógenos que são mais estudados e alvo de pesquisas do que outros, possivelmente devido às suas características e às suas interações com outros tipos de hospedeiros. Essa diferença sugere que ainda há um vasto conhecimento a ser explorado sobre os patógenos com menos artigos.

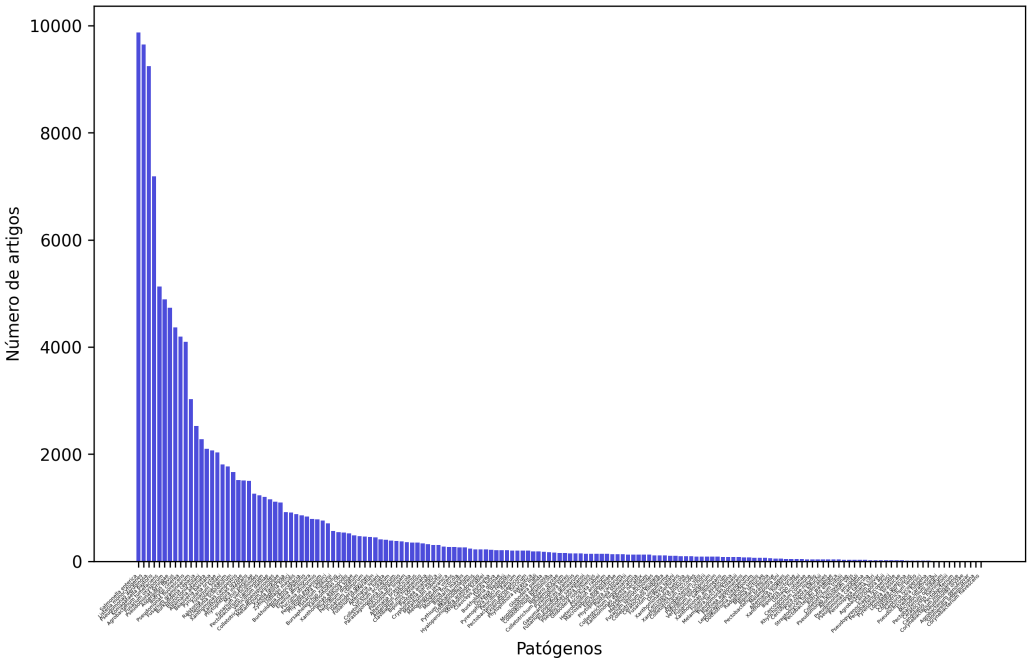

Figura 5.2: Distribuição de artigos científicos por patógeno.

5.2 Conjunto de Dados

5.2.1 Reconhecimento de Entidade Nomeadas

Utilizando a ferramenta MERpy para o reconhecimento de entidades nomeadas, identificámos um total de 275 114 menções de entidades e 190 694 entidades únicas, considerando tanto os títulos como os resumos de cada artigo. Isto representa uma média de 2.62 entidades mencionadas por artigo e uma média 1.82 de entidades únicas por artigo. Os resultados da agregação das avaliações feitas por cada anotador está disponível na tabela 5.3. As anotações de cada artigo em comum entre dois anotadores foram consideradas se ambos concordassem na anotação, onde em caso de discórdia era considerado a anotação do anotador mais especialista na área biológica. O valor da Precision indica que a maioria das entidades identificadas são realmente positivas. Por outro lado, o Recall revela que o MERpy não conseguiu identificar várias entidades presentes nos textos. Essa limitação pode ser atribuída à presença de muitas abreviaturas nos artigos, nas quais as entidades geralmente são mencionadas uma única vez pelo nome completo, enquanto nas ocorrências subsequentes são referidas apenas por abreviaturas. Dos artigos avaliados os anotadores identificaram no total 104 abreviaturas, onde nenhuma destas foi identificada pela ferramenta. No entanto 88% das entidades únicas no total dos artigos avaliados foram identificadas.

As figuras 5.3 e 5.4 representam os resumos de dois artigos diferentes extraídos do PubMed, a quando a aplicação da ferramenta MERpy, na identificação de micróbios, destacando a verde as entidades encontradas e a vermelho as entidades por encontrar. No primeiro resumo a ferramenta identificou 3 entidades, referentes a 2 micróbios diferentes, encontrando todas as entidades

Tabela 5.3: Resultados da avaliação manual da ferramenta MERpy.

Precision	Recall	F1-Score
0.938	0.578	0.716

presentes no texto, contrastando com o segundo resumo onde apenas identificou 2 entidades de 6. Contudo, é possível observar que as entidades não encontradas no segundo resumo, tratam-se de abreviaturas do micróbio *Botrytis cinerea*. Os micróbios identificados neste processo foram guardados na base de dados, permitindo a futura criação do conjunto de dados.

"**Botrytis cinerea** is one of the most serious post-harvest pathogens of fruits and vegetables. Volatiles generated by **Bacillus subtilis** JA significantly inhibited both spore germination and elongation of germ tubes in **Botrytis cinerea** using a two-compartment agar-plate assay. The volatiles caused protoplasm retraction from the hyphal tips to the spores."

Figura 5.3: Resumo do artigo Chen et al. (2008)[19], destacando a verde as entidades identificadas pela ferramenta MERpy.

"Gray mold caused by **Botrytis cinerea** is detrimental to plants and fruits. Endophytes have been shown to modify plant disease severity in functional assays. We conducted this study to investigate the endophytic strain Bacillus K1 with excellently antagonistic **B. cinerea** from the wild grape endosphere. We identified a wild grape endophytic strain K1 with high antifungal activity against **B. cinerea** both in vitro and in vivo. Combining the phylogenetic results based on 16S rDNA and genome sequencing, K1 was assigned as **Bacillus subtilis**. The in vitro results displayed that K1 and its volatile substances could significantly inhibit the mycelia growth of **B. cinerea**. Grape fruit inoculated with Bacillus K1 showed lower gray mold during treatment. The higher levels of defense-related enzymes, including peroxidase, polyphenol oxidase, and phenylalanine ammonia lyase, were induced in grapes after inoculation. Scanning electron microscopy (SEM) suggested that K1 inhibited mycelial growth via bacterial colonization and antibiosis in grapes. The gas chromatography-mass spectrometry analysis identified 33 volatiles in which dibutyl phthalate was the major compound accounting for 74.28%. Dibutyl phthalate demonstrated strong activity in suppressing the mycelia growth of **B. cinerea.**"

Figura 5.4: Excerto do resumo do artigo Li et al. (2022)[51], destacando a verde as entidades identificadas e a vermelho as não identificadas pela ferramenta MERpy.

5.2.2 Visualização do Conjunto de Dados

No total, a base de dados contém 8 476 nomes de micróbios, 255 nomes de patógenos e 104 984 artigos, com os nomes de micróbios extraídos do reconhecimento de entidades nomeadas e patógenos sem serem agrupados os sinónimos. Cada artigo está associado a pelo menos um patógeno, podendo ou não ter micróbios associados. Alguns grupos taxonómicos dos patógenos não foram adicionadas à lista por não terem sido encontrados artigos referentes a estes.

Os dados da base de dados foram usados para criar o conjunto de dados PPathoRM. Cada linha do conjunto de dados representa uma associação entre um patógeno e um micróbio com o número total de artigos onde aparecem associados. A tabela 5.4 apresenta um exemplo do conjunto de dados PPathoRM, onde é possível ver que o patógeno *Alternaria alternata* está associado a pelo menos 3 micróbios diferentes, sendo mencionado em conjunto com o micróbio *Aeromonas cavernicola* em 1 artigo, com o micróbio *Agaricus bisporus* em 3 artigos e o com o micróbio

Amorphotheca resinae em 1 artigo.

Tabela 5.4: Amostra do conjunto de dados PPathoRM

Patógeno	Micróbio	*Rating*
Alternaria alternata	Aeromonas cavernicola	1
Alternaria alternata	Agaricus bisporus	3
Alternaria alternata	Amorphotheca resinae	1
Agrobacterium tumefaciens	Aspergillus flavus	7
Agrobacterium tumefaciens	Aureimonas ureilytica	1
Agrobacterium tumefaciens	Bacillus aerophilus	1
Agrobacterium tumefaciens	Delftia acidovorans	2

A normalização dos grupos taxonómicos das entidades, tanto de micróbios como de patógenos, agrupando-as pelo nome científico, reduziu o número de patógenos de 255 para 161 e o número de micróbios de 8 476 para 7 412, no conjunto de dados. Este contém 32 726 *ratings* entre patógenos e micróbios, pelo que a esparsidade de *ratings* na matriz patógeno/micróbio é de 97.26% à semelhança de outros conjuntos de dados utilizados em sistemas de recomendação como é o caso do MovieLens100K [35] com esparsidade de 93.7%. Como referido na secção 4.5.1, criámos um conjunto de dados filtrado derivado do PPathoRM, o PPathoRM20, contendo apenas todos os patógenos que tenham pelo menos 21 micróbios associados. PPathoRM20 é constituído por 138 patógenos, 7 391 micróbios e 32 480 *ratings*. A tabela 5.5 mostra as estatísticas dos conjuntos de dados PPathoRM e PPathoRM20.

Tabela 5.5: Comparação entre os diferentes conjuntos de dados criados para Patógenos (número de patógenos), Micróbios (número de micróbios), Class (número de *ratings*), Esparsidade, MinClass (*rating* mínimo) e MaxClass (*rating* máximo).

	Patógenos	**Micróbios**	**Class**	**Esparsidade**	**MinClass**	**MaxClass**
PPathoRM	161	7 412	32 726	97.26%	1	1 446
PPathoRM20	138	7 391	32 480	96.82%	1	1 446

A Figura 5.5 representa a distribuição dos *ratings* por itens, evidenciando que o PPathoRM segue a típica distribuição de *Long Tail*. Este conceito descreve uma distribuição em que um pequeno número de itens é muito popular, enquanto os restantes itens, embora menos conhecidos, representam uma parte maioritária do total. Isto demonstra que uma pequena quantidade de micróbios foi associada a diferentes patógenos, enquanto a maioria tem associações mais esparsas, indicando que estes micróbios mais populares, por provavelmente serem mais conhecidos e estudados, podem ter sido alvo de múltiplas tentativas de associação com diferentes patógenos, na expectativa de identificar possíveis interações relevantes.

A figura 5.6 detalha o número de *ratings* por cada valor de *rating*, usando uma escala logarítmica para o número de *ratings* para melhor compreensão. O *rating* mínimo é 1 representando 61% de todos os *ratings* e o *rating* máximo 1 446, significando que há um único patógeno referen-

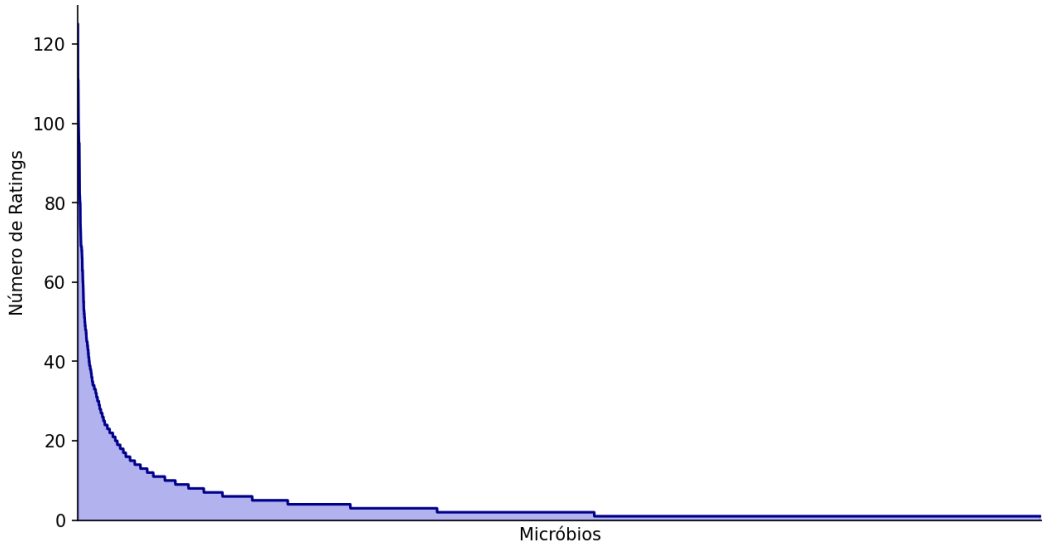

Figura 5.5: Histograma da distribuição de *ratings* por micróbio em PPathoRM

ciado 1 446 vezes em conjunto com um micróbio, correspondendo à associação entre *Salmonella enterica* e *Escherichia coli* respetivamente, onde o seguinte é 1 218.

A existência de *ratings* tão elevados sugere que há patógenos e micróbios que foram altamente estudados em conjunto. Como mencionado na secção 2.1, os patógenos são microrganismos, tal como os micróbios, diferenciando-se principalmente na forma como interagem com o hospedeiro. Os patógenos causam tipicamente doenças nos hospedeiros, mas os seus efeitos podem variar consoante o tipo de hospedeiro. Este facto pode justificar os elevados valores de *rating* atribuídos a certos pares patógeno-micróbio.

Quando um artigo contém mais de um agente tipicamente patogénico, estes também aparecem na lista de micróbios, uma vez que não podemos determinar com o processo feito até aqui se a interação entre as entidades é de cooperação ou antagonismo. Assim, é possível que duas associações com um *rating* alto correspondam a dois micróbios com potencial patogénico. Essa limitação, referente à dificuldade em determinar as interações entre as entidades, leva-nos à etapa seguinte, na extração das relações entre as entidades, com o objetivo de identificar e diferenciar essas mesmas associações.

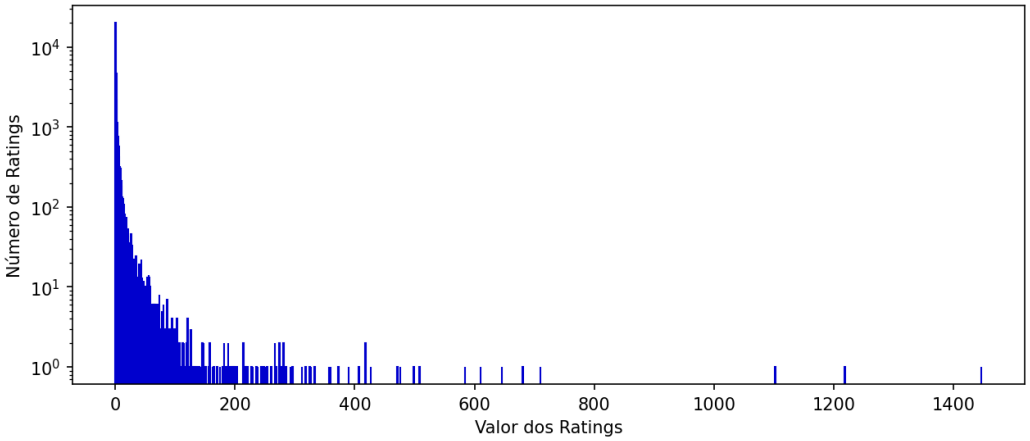

Figura 5.6: Distribuição dos valores dos *ratings* em PPathoRM

5.3 Grafo de Conhecimento

5.3.1 Extração da Relação

Na extração de relações, normalizamos as relações obtidas pelo modelo, cujo *output* deveria ser apenas uma das *tags* definidas. No entanto, nem sempre o resultado correspondia ao solicitado, divagando no tipo de resposta dada. As relações foram mapeadas manualmente, existindo algumas que era possível inferir o tipo de relação que o modelo pretendia destacar entre as duas entidades, procedendo-se à sua alteração para a respetiva *tag*. As restantes associações que não iam de encontro a nenhuma das *tags* foram substituídas por '*Unknown*'.

O uso deste tipo de modelo para a extração de relações não tem como objetivo principal avaliar o seu desempenho na identificação de relações entre micróbios e patógenos. Em vez disso, serve para inferir relações na ausência de modelos específicos para essa tarefa. Aproveita-se, assim, a capacidade dos Modelos de Linguagem de Grande Escala de generalizar com base no contexto, conforme mencionado na secção 2.3.2.

A tabela 5.6 apresenta a contagem de cada tipo de relação. A relação antagonista, que é o tipo de relação que pretendemos descobrir, representa 40% das associações encontradas. Apenas 2% das associações são agonistas. Cerca de 53% das associações o modelo não conseguiu encontrar qualquer tipo de relação entre as duas entidades.

Tabela 5.6: Contagem dos diferentes tipos de relações

Tipo de Relação	Contagem
Antagonist	12 926
Agonist	634
NoRelation	17 260
Unknown	1 906

Apesar do modelo utilizado não ser construído especificamente para encontrar este tipo de

relações entre micróbios e patógenos, a partir dos resultados consideramos é um bom passo inicial para perceber que a referência de duas entidades num artigo, não evidenciam que exista obrigatoriamente um relação entre estas.

5.3.2 Visualização do Grafo de Conhecimento

A base de dados do Neo4j contém o grafo de conhecimento com as diferentes entidades e relações extraídas até aqui. Cada nó e aresta, para além das propriedades descritas na secção 4.4 contêm dois identificadores únicos, um global e outro interno, característicos do Neo4j, permitindo diferenciar e referenciar de forma eficaz cada elemento da base de dados. A base de dados ocupa um total de 31,8 megabytes em disco.

A figura 5.7 apresenta a visualização de uma amostra do grafo conhecimento, na base de dados do Neo4j, representando o patógeno *Botrytis cinerea* e diversas das suas relações com micróbios. Por exemplo, na imagem observa-se que o micróbio *Bacillus subtilis* contém uma relação de Antagonismo com o patógeno em referência. Este tipo de análise é facilitado pelas capacidades interativas do Neo4j, que permite explorar e analisar o grafo de conhecimento de forma mais profunda e detalhada. As figuras A.1 e A.2 representam diferentes visualizações do grafo.

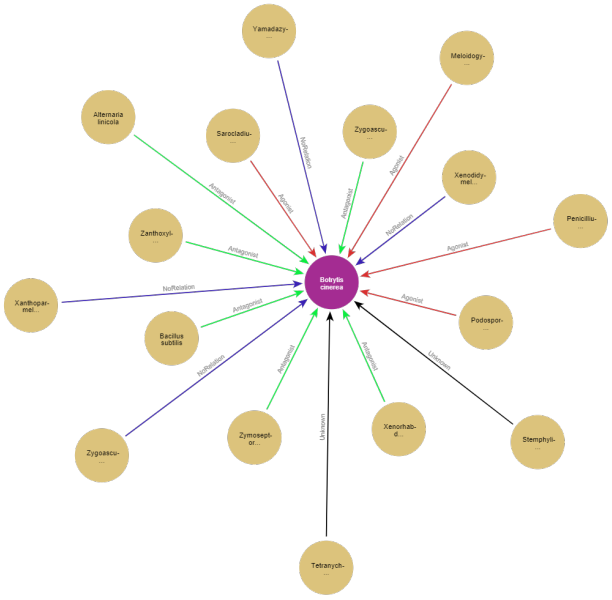

Figura 5.7: Visualização de uma amostra do grafo de conhecimento na base de dados do Neo4j, para o patógeno *Botrytis cinerea* com algumas das suas relações com micróbios.

5.4 Sistema de Recomendação

5.4.1 Validação do Conjunto de Dados

A validação do nosso conjunto de dados é importante para perceber se este é aplicável em sistemas de recomendação. A figura 5.8 mostra os resultados obtidos ao aplicar dois algoritmos de

recomendação diferentes, ALS e KNN, nos conjuntos de dados PPathoRM e PPathoRM20, para a Precision, Recall, F1-Score, MRR e nDCG, para diferentes números de itens recomendados.

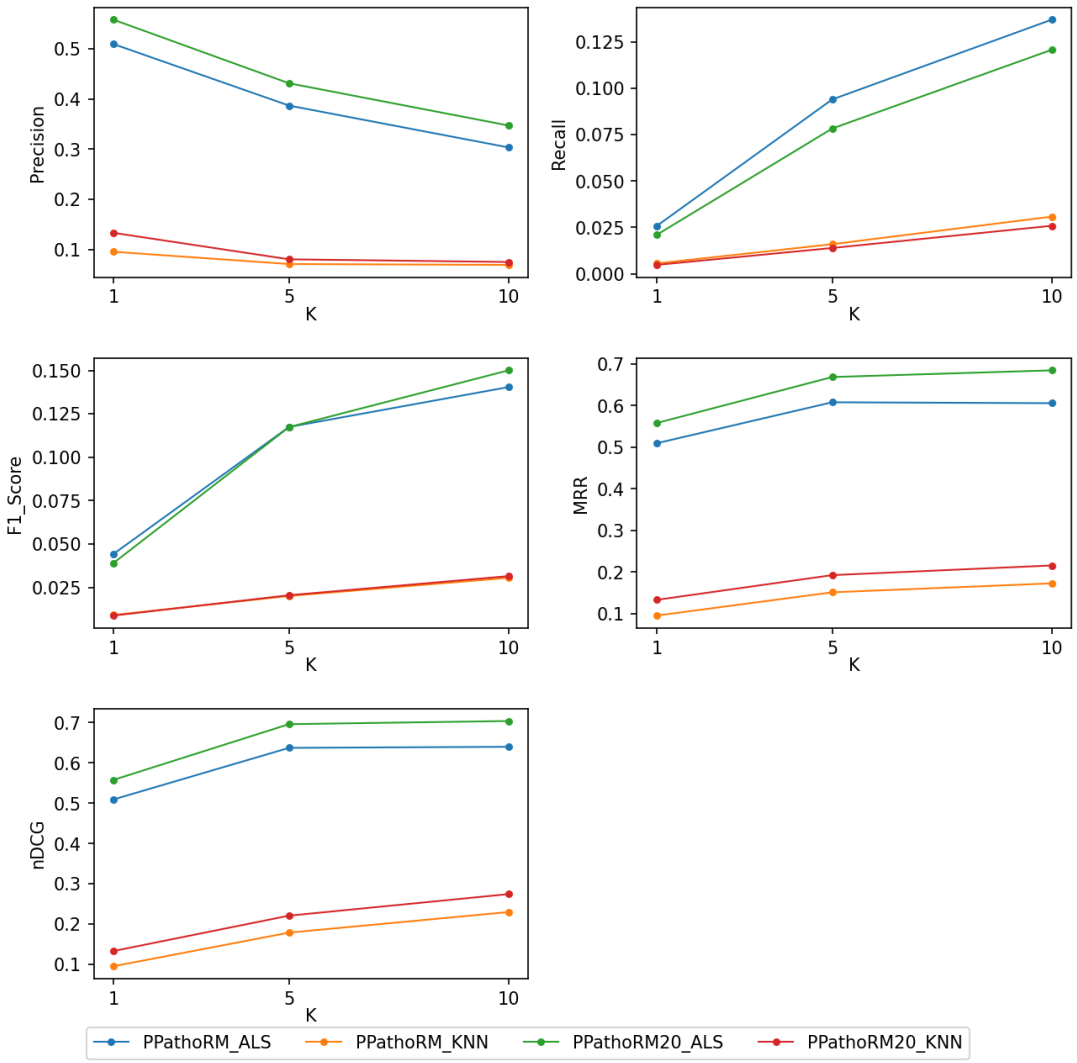

Figura 5.8: Resultados da recomendação para PPathoRM e PPathoRM20, com os algoritmos ALS e KNN, para as métricas Precision, Recall, F1-Score, MRR e nDCG, em função de K (número de itens recomendados)

Analisando a figura 5.8, o algoritmo ALS apresenta um desempenho consistentemente superior ao algoritmo KNN em todas as métricas de avaliação, independentemente do conjunto de dados utilizado. Isto é justificado, dado que o ALS, sendo uma técnica de fatorização de matrizes, é bastante eficiente ao lidar com dados altamente esparsos e implícitos, como é o caso de PPathoRM e PPathoRM20.

Ambos os algoritmos seguem tendências semelhantes à medida que o número de itens recomendados (K) aumenta. A Precision é maior quando apenas 1 item é recomendado, mas diminui progressivamente com o aumento de K. Isto ocorre porque, à medida que mais itens são recomendados, há uma maior probabilidade de incluir itens irrelevantes na lista. Por outro lado, o Recall segue a tendência oposta, aumentando à medida que mais itens são recomendados, uma vez que há

mais chances de capturar todos os itens relevantes. No entanto, os valores de Recall apresentados são relativamente baixos, variando de 0.004 a 0.137, refletindo a dificuldade em capturar todas as possíveis interações relevantes no conjunto de dados. Já os valores de Precision são significativamente mais elevados, com o melhor resultado a ser 0.56 para o ALS no conjunto PPathoRM20.

As métricas de MRR e nDCG apresentam bons resultados para ALS, especialmente quando 5 ou mais itens são recomendados, o que indica que a lista de itens apresenta uma boa qualidade na ordem de recomendação dos itens.

Em termos de comparação entre os dois conjuntos de dados, o PPathoRM destaca-se por apresentar melhor Recall, independentemente do número de itens recomendados. Contudo, para as restantes métricas, os resultados do PPathoRM20 são superiores, especialmente em Precision.

Estes resultados estão de acordo com os resultados apresentados noutros estudos [10], demonstrando a viabilidade desde conjunto de dados para testar sistemas de recomendação na área das doenças das plantas. No entanto, uma das limitações é o seu tamanho, dado que os conjuntos de dados utilizados em sistemas de recomendação tendem a ser maiores, o que pode influenciar ligeiramente os resultados. Além disso, o conjunto de dados apenas permite recomendar micróbios que de alguma forma tenham tido uma interação com o patógeno, não permitindo dizer se é uma relação antagonista ou agonista.

5.4.2 Recomendação de Antagonistas

As relações fornecem uma informação extra e bastante relevante. A partir delas é possível perceber se um micróbio de facto é capaz de inibir ou não um patógeno. No conjunto de dados PPathoRM apenas constam as diversas associações presentes em diferentes artigos, contudo não existe nenhuma informação acerca do tipo de interação. Na secção 5.4.1 este conjunto foi validado considerando todo o tipo de interações como positivas. Para existir uma comparação entre as recomendações efetuadas a partir desse conjunto de dados e do grafo de conhecimento, é necessário reduzir os itens relevantes apenas aos que são realmente antagonistas e não a qualquer associação. Usamos essa abordagem para definir um modelo base e perceber se as recomendações a partir do grafo de conhecimento melhoram a qualidade das recomendações.

De forma a percebermos o que está a ser recomendado a figura 5.9 representa a recomendação do top 10 de micróbios com relações antagonistas relativamente ao patógeno 4286 (*Colletotrichum higginsianum*), usando DistMult. Para esta, a lista de itens recomendados é [991, 864, 185, 136, 140, 435, 1118, 843, 837, 433] correspondendo respetivamente aos nomes de micróbios [Sclerotinia sclerotiorum, Pseudomonas syringae, Botrytis cinerea, Bacillus amyloliquefaciens, Bacillus cereus, Fusarium verticillioides, Trichoderma harzianum, Pseudomonas fluorescens, Pseudomonas aeruginosa, Fusarium solani]. Os nomes destacados pertencem à lista de itens relevantes referentes a este patógeno, sendo os corretamente recomendados. Para este patógeno a Precision@10 é de 0.2 e o Recall@10 de 0.18, sendo que a lista de itens relevantes no conjunto de teste contém 11 micróbios.

A figura 5.10 apresenta a comparação dos resultados obtidos, pela aplicação do algoritmo

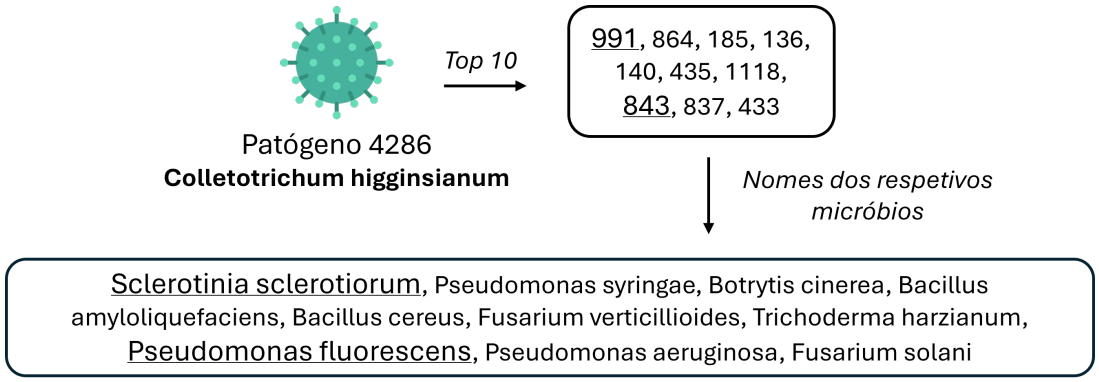

Figura 5.9: Exemplo de recomendação do top 10 de micróbios para o patógeno *Colletotrichum higginsianum*, com DistMult, usando PPathoKG, destacando os itens relevantes.

ALS no conjunto de dados PPathoRM e pelo uso dos embeddings DistMult em PPathoKG para recomendação de micróbios. O modelo de ALS serve como modelo base, considerando que o conjunto de dados PPathoRM não contém qualquer tipo de distinção entre associações e apenas recomenda com base nos *ratings*. Já os embeddings produzidos pelo DistMult são treinados com base nos diferentes tipos de relações, ignorando os *ratings*.

DistMult apresenta melhores resultados em praticamente todas as métricas, quando o número de itens recomendados é de 1, 5 e 10. O valor da Precision é o que mais se destaca com uma maior diferença, sendo de 0.256 para DistMult e 0.189 para o ALS, quando apenas 1 item é recomendado. O Recall apresenta valores aproximados, onde o ALS apresenta o único valor superior ao DistMult, quando se recomenda 10 itens. No entanto este apresenta valores demasiados baixos, significando que a maioria dos itens relevantes não estão a ser recomendados. Isto é facilmente justificável quando analisamos o número de itens relevantes de cada patógeno. Diversos patógenos apresentam um número elevado de micróbios antagonistas, o que se traduz em muitos itens relevantes. Ao limitarmos o nosso número de itens recomendados a 1, 5 ou 10, não damos margem a que muitos dos itens relevantes sejam capturados, mesmo que todos os recomendados sejam relevantes.

Os resultados demonstram que o PPathoKG, ao integrar *ratings* com relações, fornece melhores resultados que as recomendações produzidas a partir de PPathoRM. Embora o ALS usado em PPathoRM seja amplamente reconhecido como um dos algoritmos mais eficazes para recomendação em conjuntos de dados de *feedback* implícito, o uso de relações pelo DistMult em PPathoKG mostrou-se superior. Como a divisão dos conjuntos de dados foi a mesma, esses resultados são diretamente comparáveis, comprovando que a inclusão de relações melhora significativamente a qualidade das recomendações. Além disso, ao considerar algoritmos mais avançados, com a integração de embeddings do grafo de conhecimento com outras técnicas, como alguns dos estudos apresentados na secção 3.2, o uso das relações oferece ainda mais potencial para futuras melhorias na qualidade do sistema de recomendação.

Focando no principal objetivo do sistema de recomendação, que é identificar novos micróbios capazes de inibir patógenos específicos, a métrica mais importante a considerar é a Precision, ou seja, garantir que as recomendações feitas são efetivamente de micróbios com capacidades

antagonistas, pois cada micróbio recomendado precisa ser testado experimentalmente.

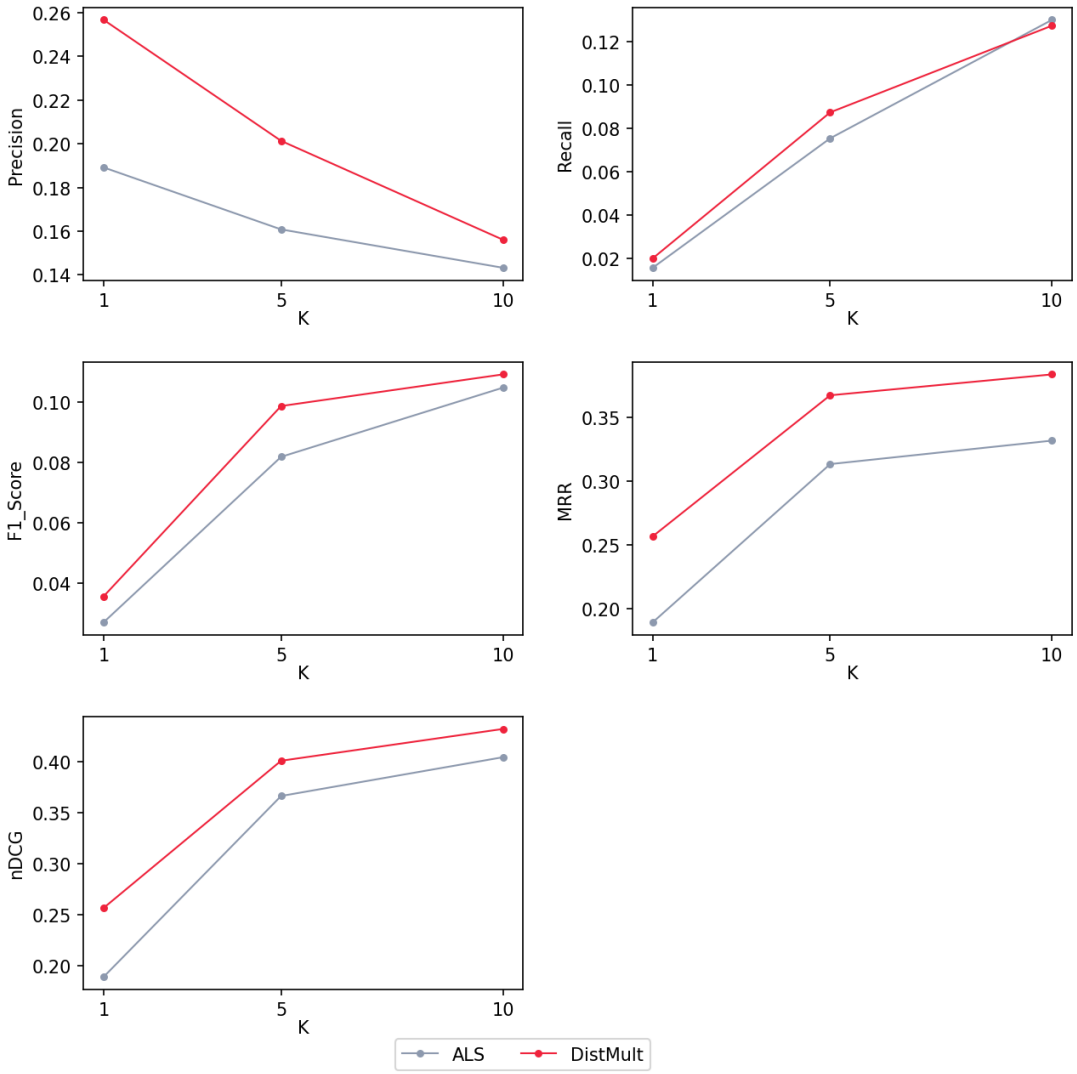

Figura 5.10: Comparação dos resultados obtidos entre o uso de PPathoRM com ALS e PPathoKG com DistMult, para as métricas Precision, Recall, F1-Score, MRR e nDCG, em função de K (número de itens recomendados)

Capítulo 6

Conclusão

O acrescido número de doenças biológicas ao longo dos anos a afetar a saúde das plantas e a produção alimentar enfatiza a importância de encontrar soluções inovadoras e eficazes no combate a patógenos. As interações entre micróbios e patógenos apresentam-se como uma área de crescente interesse científico, oferecendo um caminho promissor para o desenvolvimento de produtos de controlo biológico naturais que possam conter os efeitos adversos destas doenças.

Neste contexto, o presente trabalho explora numa primeira fase a criação de um conjunto de dados, PPathoRM, de *feedback* implícito a partir da literatura científica, usando a metodologia LIBRETTI, com as diversas associações entre micróbios e patógenos de plantas, com o objetivo de criar um sistema de recomendação de micróbios para patógenos de plantas.

Utilizamos o MERpy na identificação de micróbios presentes na literatura, realizando uma avaliação manual. As anotações encontradas são maioritariamente verdadeiras positivas, mas diversas anotações ficaram por serem encontradas, devido ao uso de muitas abreviaturas na literatura científica. Apesar disso a maior parte dos micróbios foram identificados, por existir geralmente sempre uma anotação com o nome completo da entidade antes de usada outras abreviaturas.

Criámos uma versão diferente do conjunto de dados, PPathoRM20, removendo todos os patógenos com menos de 20 micróbios associados. Esta abordagem é avaliada aplicando o algoritmo ALS e KNN nos conjuntos de dados PPathoRM e PPathoRM20, através de diversas métricas. O ALS destacou-se em todas as métricas, para qualquer um dos conjuntos de dados, revelando-se como já esperado, um dos melhores algoritmos para recomendações em conjuntos de dados de *feedback* implícito.

A segunda fase pretende melhorar a abordagem com a criação de um grafo de conhecimento, PPathoKG, que incorpora adicionalmente as relações entre os micróbios e patógenos, extraídas usando o Modelo de Linguagem de Grande Escala Mixtral-8x7B-Instruct-v0.1. Procuramos verificar se a adição das relações influenciava a qualidade das recomendações. Para isso, utilizámos o algoritmo ALS, aplicado no conjunto de dados PPathoRM, como modelo base. Em seguida, implementámos o modelo DistMult no grafo de conhecimento para gerar embeddings que capturam as interações entre as entidades e, assim, identificar potenciais relações antagonistas. A avaliação comparativa entre os dois algoritmos, ALS e DistMult, mostrou que o ALS, embora se destaque na recomendação de micróbios baseados apenas em *feedback* implícito, foi superado pelo DistMult

para qualquer uma das métricas, quando consideramos o tipo de relação existente entre entidades.

O grafo de conhecimento criado oferece uma contribuição significativa ao conter todas as relações entre micróbios e patógenos presentes na literatura científica extraída para este estudo entre 2003 e 2024. Este grafo não só responde à necessidade da procura imediata de produtos biológicos para tratamento de patógenos em plantas nos campos agrícolas, mas também serve de base para um sistema de recomendação que otimiza o processo de seleção de micróbios com capacidades antagonistas. A redução do número de micróbios que precisam ser testados contra patógenos, acelera a identificação de produtos biológicos eficazes para o controlo de doenças de plantas, tornando o processo mais eficiente e eficaz.

6.1 Trabalho Futuro

Este trabalho abre caminho para novos estudos e desenvolvimentos. Uma das potenciais melhorias é a adição de características específicas dos micróbios, possibilitando a implementação de um algoritmo de filtragem em conteúdo com a filtragem colaborativa, criando um sistema híbrido. Outra direção poderia ser associar cada um dos patógenos a uma planta que este provoque uma doença conhecida, explorando se a interação de um micróbio com um patógeno varia dependendo do tipo de planta.

Avaliar novos métodos de recomendação a partir do PPathoKG, considerando os *ratings* presentes nas relações com as diferentes tipos de relações, ou o uso de algoritmos mais avançados que combinem embeddings do grafo de conhecimento com outras técnicas, como o KHGCN [32], podem permitir capturar interações mais complexas entre os micróbios e patógenos.

Bibliografia

[1] Kessy Abarenkov, Allan Zirk, Timo Piirmann, Raivo Pöhönen, Filipp Ivanov, R. Henrik Nilsson, and Urmas. Kõljalg. Unite general fasta release for eukaryotes 2, 2023. [Online - Accessed on 27-11-2023].

[2] Mina Abbaszade, Vahid Salari, Seyed Shahin Mousavi, Mariam Zomorodi, and Xujuan Zhou. Application of quantum natural language processing for language translation. *IEEE Access*, 9:130434–130448, 2021.

[3] Charu C Aggarwal and Charu C Aggarwal. Ensemble-based and hybrid recommender systems. *Recommender Systems: The Textbook*, pages 199–224, 2016.

[4] Nisar Ahmad and Samayveer Singh. Comparative study of disease detection in plants using machine learning and deep learning. In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, pages 54–59. IEEE, 2021.

[5] Xavier Amatriain. Beyond data: from user information to business value through personalized recommendations and consumer science. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2201–2208, 2013.

[6] Ali A Amer, Hassan I Abdalla, and Loc Nguyen. Enhancing recommendation systems performance using highly-effective similarity measures. *Knowledge-Based Systems*, 217:106842, 2021.

[7] Pradeepa Bandara, Thilini Weerasooriya, T Ruchirawya, W Nanayakkara, M Dimantha, and M Pabasara. Crop recommendation system. *International Journal of Computer Applications*, 975:8887, 2020.

[8] Miranda Barnes and Dana C Price. Endogenous viral elements in ixodid tick genomes. *Viruses*, 15(11):2201, 2023.

[9] Matilde Barón, Mónica Pineda, and María Luisa Pérez-Bueno. Picturing pathogen infection in plants. *Zeitschrift für Naturforschung C*, 71(9-10):355–368, 2016.

[10] Márcia Barros, André Moitinho, and Francisco M Couto. Using research literature to generate datasets of implicit feedback for recommending scientific items. *IEEE Access*, 7:176668–176680, 2019.

[11] Jasmin Bharadiya. A comprehensive survey of deep learning techniques natural language processing. *European Journal of Technology*, 7(1):58–66, 2023.

[12] Hernando José Bolivar-Anillo, Victoria E González-Rodríguez, Jesús M Cantoral, Darío García-Sánchez, Isidro G Collado, and Carlos Garrido. Endophytic bacteria bacillus subtilis, isolated from zea mays, as potential biocontrol agent against botrytis cinerea. *Biology*, 10(6):492, 2021.

[13] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[14] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12:331–370, 2002.

[15] Richard Ewen Campbell. *Biological control of microbial plant pathogens*. Cambridge university press, 1989.

[16] Arturo Casadevall. The pathogenic potential of a microbe. *Msphere*, 2(1):10–1128, 2017.

[17] Brad Chapman and Jeffrey Chang. Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, 20(2):15–19, 2000.

[18] Fukun Chen, Guisheng Yin, Yuxin Dong, Gesu Li, and Weiqi Zhang. Khgcn: Knowledge-enhanced recommendation with hierarchical graph capsule network. *Entropy*, 25(4):697, 2023.

[19] Hua Chen, Xiang Xiao, Jun Wang, Lijun Wu, Zhiming Zheng, and Zengliang Yu. Antagonistic effects of volatiles generated by bacillus subtilis on spore germination and hyphal growth of the plant pathogen, botrytis cinerea. *Biotechnology letters*, 30:919–923, 2008.

[20] Mahendra Choudhary, Rohit Sartandel, Anish Arun, Leena Ladge, Saroj Hiranwal, and Garima Mathur. Crop recommendation system and plant disease classification using machine learning for precision agriculture. In *Artificial Intelligence and Communication Technologies, SCRS*, pages 39–49, 2022.

[21] S Ciampelli, AE Voppel, JN De Boer, S Koops, and IEC Sommer. Combining automatic speech recognition with semantic natural language processing in schizophrenia. *Psychiatry research*, 325:115252, 2023.

[22] Francisco M Couto and Andre Lamurias. Mer: a shell script and annotation server for minimal named entity recognition and linking. *Journal of cheminformatics*, 10:1–10, 2018.

[23] Makaylee K Crone, Natalie K Boyle, Sean T Bresnahan, David J Biddinger, Rodney T Richardson, and Christina M Grozinger. More than mesolectic: Characterizing the nutritional niche of osmia cornifrons. *Ecology and Evolution*, 13(10):e10640, 2023.

[24] Pedro W Crous, Walter Gams, Joost A Stalpers, Vincent Robert, and Gerrit Stegehuis. Mycobank: an online initiative to launch mycology into the 21st century. *Studies in mycology*, 50(1):19–22, 2004.

[25] KPK Devan, B Swetha, P Uma Sruthi, and S Varshini. Crop yield prediction and fertilizer recommendation system using hybrid machine learning algorithms. In *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, pages 171–175. IEEE, 2023.

[26] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2, 2016.

[27] Mohamed T El-Saadony, Ahmed M Saad, Soliman M Soliman, Heba M Salem, Alshaymaa I Ahmed, Mohsin Mahmood, Amira M El-Tahan, Alia AM Ebrahim, Abd El-Mageed, A Taia, et al. Plant growth-promoting microorganisms as biocontrol agents of plant diseases: Mechanisms, challenges and future perspectives. *Frontiers in plant science*, 13:923880, 2022.

[28] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[29] National Center for Biotechnology Information (US). Entrez Programming Utilities Help, 2010.

[30] Ben Frederickson. Fast python collaborative filtering for implicit datasets. *URL https://github. com/benfred/implicit*, 2018.

[31] F Furtado and A Singh. Movie recommendation system using machine learning. *International journal of research in industrial engineering*, 9(1):84–98, 2020.

[32] David Geleta, Andriy Nikolov, Gavin Edwards, Anna Gogleva, Richard Jackson, Erik Jansson, Andrej Lamov, Sebastian Nilsson, Marina Pettersson, Vladimir Poroshin, et al. Biological insights knowledge graph: an integrated knowledge graph to support drug development. *Biorxiv*, pages 2021–10, 2021.

[33] José Guia, Valéria Gonçalves Soares, and Jorge Bernardino. Graph databases: Neo4j analysis. In *ICEIS (1)*, pages 351–356, 2017.

[34] Ankit Gupta, Rasna Gupta, and Ram Lakhan Singh. Microbes and environment. *Principles and applications of environmental biotechnology for a sustainable future*, pages 43–84, 2017.

[35] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[36] Yu-An Huang, Zhu-Hong You, Xing Chen, Zhi-An Huang, Shanwen Zhang, and Gui-Ying Yan. Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. *Journal of translational medicine*, 15(1):1–11, 2017.

[37] Folasade Olubusola Isinkaye, Yetunde O Folajimi, and Bolande Adefowoke Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3):261–273, 2015.

[38] Sarika Jain, Anjali Grover, Praveen Singh Thakur, and Sourabh Kumar Choudhary. Trends, problems and solutions of recommender system. In *International conference on computing, communication & automation*, pages 955–958. IEEE, 2015.

[39] Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer, 2022.

[40] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. Comparison of implicit and explicit feedback from an online music recommendation service. In *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems*, pages 47–51, 2010.

[41] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[42] Anu Jose, S Nandagopalan, Vidya Ubalanka, and Dhanya Viswanath. Detection and classification of nutrient deficiencies in plants using machine learning. In *Journal of Physics: Conference Series*, volume 1850, page 012050. IOP Publishing, 2021.

[43] Monisha Kanakaraj and Ram Mohana Reddy Guddeti. Nlp based sentiment analysis on twitter data using ensemble classifiers. In *2015 3Rd international conference on signal processing, communication and networking (ICSCN)*, pages 1–5. IEEE, 2015.

[44] Nikitha Karkera, Sathwik Acharya, and Sucheendra K Palaniappan. Leveraging pre-trained language models for mining microbiome-disease relationships. *BMC bioinformatics*, 24(1):290, 2023.

[45] Seyed Mehran Kazemi and David Poole. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems*, 31, 2018.

[46] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

[47] Vinod Kumar and Piyush Kumar. Pesticides in agriculture and environment: Impacts on human health. *Contaminants in agriculture and environment: health risks and remediation*, 1:76–95, 2019.

[48] Thanh Le, Nam Le, and Bac Le. Knowledge graph embedding by relational rotation and complex convolution for link prediction. *Expert Systems with Applications*, 214:119122, 2023.

[49] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[50] Benjamin Yee Shing Li, Lam Fat Yeung, and Genke Yang. Pathogen host interaction prediction via matrix factorization. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 357–362. IEEE, 2014.

[51] Peiqian Li, Baozhen Feng, Zhen Yao, Bohui Wei, Yanfei Zhao, and Shouguo Shi. Antifungal activity of endophytic bacillus k1 against botrytis cinerea. *Frontiers in Microbiology*, 13:935675, 2022.

[52] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, volume 380, pages 2739–2745, 2020.

[53] Yue Liu, Shu-Lin Wang, Jun-Feng Zhang, Wei Zhang, and Wen Li. A neural collaborative filtering method for identifying mirna-disease associations. *Neurocomputing*, 422:176–185, 2021.

[54] Yahui Long, Min Wu, Chee Keong Kwoh, Jiawei Luo, and Xiaoli Li. Predicting human microbe–drug associations via graph convolutional network with conditional random field. *Bioinformatics*, 36(19):4918–4927, 2020.

[55] John A Lucas, Nichola J Hawkins, and Bart A Fraaije. The evolution of fungicide resistance. *Advances in applied microbiology*, 90:29–92, 2015.

[56] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282, 2022.

[57] Tareq B Malas, Wytze J Vlietstra, Roman Kudrin, Sergey Starikov, Mohammed Charrout, Marco Roos, Dorien JM Peters, Jan A Kors, Rein Vos, Peter AC ‘t Hoen, et al. Drug prioritization using the semantic properties of a knowledge graph. *Scientific reports*, 9(1):6281, 2019.

[58] Gesmond George Manuval, Thomas T George, Bilha P Aby, Mohith Mathew, Ayush Sarath Chandran, and N Jayapandian. Machine learning based candidate recommendation system using bayesian model. In *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1172–1178. IEEE, 2023.

[59] Michael McLaren. Silva 138.1 prokaryotic ssu taxonomic training data formatted for dada2, March 2021. [Online - Accessed on 27-11-2023].

[60] Angela Medvedeva, Hamid Teimouri, and Anatoly B Kolomeisky. Predicting antimicrobial activity for untested peptide-based drugs using collaborative filtering and link prediction. *Journal of Chemical Information and Modeling*, 63(12):3697–3704, 2023.

[61] Kurt Mendgen and Matthias Hahn. Plant infection and the establishment of fungal biotrophy. *Trends in plant science*, 7(8):352–356, 2002.

[62] Miftahul Jannat Mokarrama and Mohammad Shamsul Arefin. Rsf: A recommendation system for farmers. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 843–850. IEEE, 2017.

[63] Jéssica Monteiro, Filipe Sá, and Jorge Bernardino. Experimental evaluation of graph databases: Janusgraph, nebula graph, neo4j, and tigergraph. *Applied Sciences*, 13(9):5770, 2023.

[64] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1):1–39, 2021.

[65] National Center for Biotechnology Information (NCBI). Ncbi taxonomy database, 2024.

[66] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. *Queue*, 17(2):48–75, apr 2019.

[67] Kalyani Pakhale. Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. *arXiv preprint arXiv:2309.14084*, 2023.

[68] Yesol Park, Joohong Lee, Heesang Moon, Yong Suk Choi, and Mina Rho. Discovering microbe-disease associations from the literature using a hierarchical long short-term memory network and an ensemble parser model. *Scientific reports*, 11(1):4490, 2021.

[69] Aidan C Parte, Joaquim Sardà Carbasse, Jan P Meier-Kolthoff, Lorenz C Reimer, and Markus Göker. List of prokaryotic names with standing in nomenclature (lpsn) moves to the dsmz. *International journal of systematic and evolutionary microbiology*, 70(11):5607–5612, 2020. [Online - Accessed on 27-11-2023].

[70] Nadeesha Perera, Thi Thuy Linh Nguyen, Matthias Dehmer, and Frank Emmert-Streib. Comparison of text mining models for food and dietary constituent named-entity recognition. *Machine Learning and Knowledge Extraction*, 4(1):254–275, 2022.

[71] Mrs.'s. Pudumalar, M. Suriya, M. P. Ramanujam, and Dr. S. Muthuramalingam. Pesticide recommendation system for cotton crop diseases due to the climatic changes. 2018.

[72] S Pudumalar, E Ramanujam, R Harine Rajashree, C Kavya, T Kiruthika, and J Nisha. Crop recommendation system for precision agriculture. In *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pages 32–36. IEEE, 2017.

[73] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596, 2012.

[74] V Robert, G Stegehuis, and J Stalpers. The mycobank engine and related databases. *www. mycobank. org*, 2005. [Online - Accessed on 27-11-2023].

[75] Vincent Robert, Duong Vu, Ammar Ben Hadj Amor, Nathalie van de Wiele, Carlo Brouwer, Bernard Jabas, Szaniszlo Szoke, Ahmed Dridi, Maher Triki, Samy ben Daoud, et al. Mycobank gearing up for new horizons. *IMA fungus*, 4:371–379, 2013.

[76] Scott Sherrill-Mix. *taxonomizr: Functions to work with NCBI Accessions and Taxonomy*, 2024. Accessed on 1-12-2023.

[77] Brajesh K Singh, Manuel Delgado-Baquerizo, Eleonora Egidi, Emilio Guirado, Jan E Leach, Hongwei Liu, and Pankaj Trivedi. Climate change impacts on plant pathogens, food security and paths forward. *Nature Reviews Microbiology*, pages 1–17, 2023.

[78] Sonit Singh. Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*, 2018.

[79] Taranjeet Singh, Krishna Kumar, and SS Bedi. A review on artificial intelligence techniques for disease recognition in plants. In *IOP Conference Series: Materials Science and Engineering*, volume 1022, page 012032. IOP Publishing, 2021.

[80] Shalini Christabel Stephen, Hong Xie, and Shri Rai. Measures of similarity in memory-based collaborative filtering recommender system: A comparison. In *Proceedings of the 4th multidisciplinary international social networks conference*, pages 1–8, 2017.

[81] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

[82] Gytė Tamašauskaitė and Paul Groth. Defining a knowledge graph development process through a systematic review. *ACM Transactions on Software Engineering and Methodology*, 32(1):1–40, 2023.

[83] Anne E Thessen, Laurel Cooper, Tyson L Swetnam, Harshad Hegde, Justin Reese, Justin Elser, and Pankaj Jaiswal. Using knowledge graphs to infer gene expression in plants. *Frontiers in Artificial Intelligence*, 6:1201002, 2023.

[84] Vinoy Koshy Thomas, Jusbin Mathew, Nivin Emmanuel, and Seban V Mathew. A plant identification and recommendation system. *International Research Journal of Engineering and Technology (IRJET)*, 6(06):2395–0056, 2019.

[85] Tanmay Thorat, BK Patle, and Sunil Kumar Kashyap. Intelligent insecticide and fertilizer recommendation system based on tpf-cnn for smart farming. *Smart Agricultural Technology*, 3:100114, 2023.

[86] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

[87] Martin Urban, Alayne Cuzick, James Seager, Valerie Wood, Kim Rutherford, Shilpa Yagwakote Venkatesh, Jashobanta Sahu, S Vijaylakshmi Iyer, Lokanath Khamari, Nishadi De Silva, et al. Phi-base in 2022: a multi-species phenotype database for pathogen–host interactions. *Nucleic Acids Research*, 50(D1):D837–D847, 2022.

[88] Aleksandar N Veljković, Yuriy L Orlov, and Nenad S Mitić. Biograph: Data model for linking and querying diverse biological metadata. *International Journal of Molecular Sciences*, 24(8):6954, 2023.

[89] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014.

[90] Somin Wadhwa, Silvio Amir, and Byron C Wallace. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access, 2023.

[91] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*, 2023.

[92] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958, 2019.

[93] Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat-Seng Chua. Learning intents behind interactions with knowledge graph for recommendation. In *Proceedings of the web conference 2021*, pages 878–887, 2021.

[94] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

[95] Wei Wei, Sen Zhao, and Ding Zou. Recommendation system: A survey and new perspectives. *World Scientific Annual Review of Artificial Intelligence*, 1:2330001, 2023.

[96] Justus Wesseler. The eu's farm-to-fork strategy: An assessment from the perspective of agricultural economics. *Applied Economic Perspectives and Policy*, 44(4):1826–1843, 2022.

[97] Hans Christian Wittich, Marco Seeland, Jana Wäldchen, Michael Rzanny, and Patrick Mäder. Recommending plant taxa for supporting on-site species identification. *BMC bioinformatics*, 19(1):1–17, 2018.

[98] Chengkun Wu, Xinyi Xiao, Canqun Yang, JinXiang Chen, Jiacai Yi, and Yanlong Qiu. Mining microbe–disease interactions from literature via a transfer learning model. *BMC bioinformatics*, 22:1–15, 2021.

[99] Da Xu, Hanxiao Xu, Yusen Zhang, and Rui Gao. Novel collaborative weighted nonnegative matrix factorization improves prediction of disease-associated human microbes. *Frontiers in Microbiology*, 13:834982, 2022.

[100] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[101] Deqing Yang, Zikai Guo, Ziyi Wang, Juyang Jiang, Yanghua Xiao, and Wei Wang. A knowledge-enhanced deep recommendation framework incorporating gan-based models. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1368–1373. IEEE, 2018.

[102] Pelin Yilmaz, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. The silva and "all-species living tree project (ltp)" taxonomic frameworks. *Nucleic acids research*, 42(D1):D643–D648, 2014.

[103] Eva Zangerle and Christine Bauer. Evaluating recommender systems: survey and framework. *ACM Computing Surveys*, 55(8):1–38, 2022.

[104] Jin-Cheng Zhang, Azlan Mohd Zain, Kai-Qing Zhou, Xi Chen, and Ren-Min Zhang. A review of recommender systems based on knowledge graph embedding. *Expert Systems with Applications*, page 123876, 2024.

[105] Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song, and Dik Lun Lee. Meta-graph based recommendation fusion over heterogeneous information networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 635–644, 2017.

[106] Tranos Zuva, Sunday O Ojo, Seleman Ngwira, Keneilwe Zuva, et al. A survey of recommender systems techniques, challenges and evaluation metrics. *International Journal of Emerging Technology and Advanced Engineering*, 2(11):382–386, 2012.

Apêndice A

Visualizações do grafo

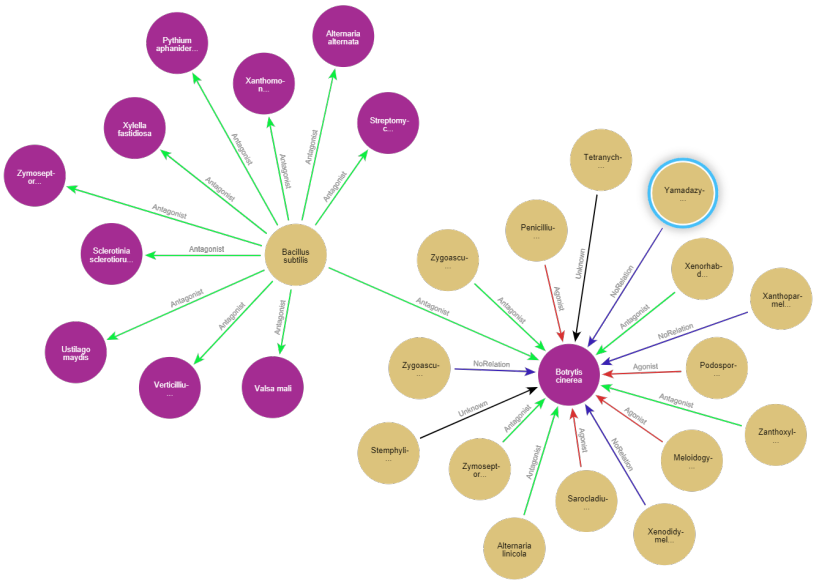

Figura A.1: Visualização de uma amostra do grafo do conhecimento na base de dados do Neo4j, para o patógeno *Botrytis cinerea* com algumas das suas relações com micróbios, e para o micróbio *Bacillus subtilis* associado a este patógeno e outros de que é antagonista.

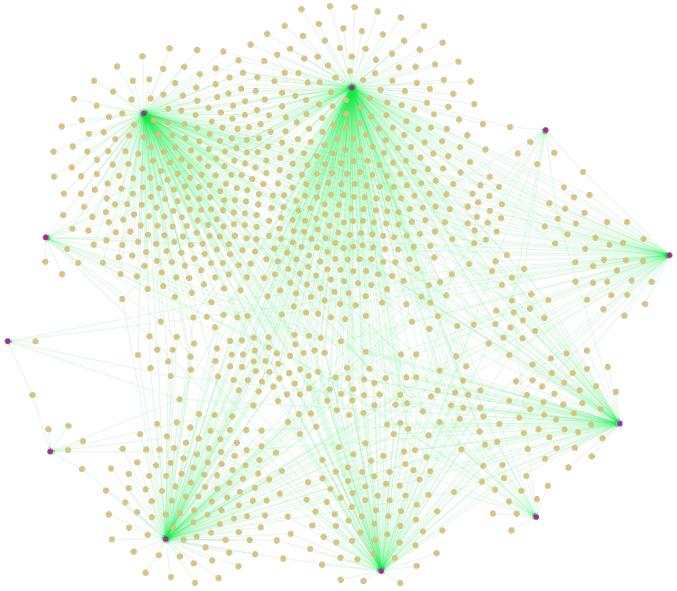

Figura A.2: Visualização de uma amostra do grafo do conhecimento na base de dados do Neo4j, para os 11 primeiros patógenos, por ordem alfabética, e os micróbios com relações antagonistas com estes.