Universidade do Minho
Escola de Engenharia

Cátia Vieira e Mendonça

Sistemas de Recomendação Científica: Incorporação da Reputação Científica nos seus Algoritmos

março de 2022

Sistemas de Recomendação Científica: Incorporação da Reputação Científica nos seus Algoritmos

Cátia Vieira e Mendonça

UMinho | 202X

Universidade do Minho
Escola de Engenharia

Cátia Vieira e Mendonça
(pg41145)

Sistemas de Recomendação Científica: Incorporação da Reputação Científica nos seus Algoritmos

Dissertação de Mestrado
Mestrado em Sistemas de Informação

Trabalho efetuado sob a orientação de
Professor Doutor Miguel Abrunhosa Brito
Professora Doutora Ana Alice Baptista

março de 2022

DIREITOS DE AUTOR

AGRADECIMENTOS

A realização da presente dissertação reúne o contributo de várias pessoas, às quais estou extremamente grata. Sem elas, este trabalho não seria possível.

Em primeiro lugar, não posso deixar de agradecer aos meus orientadores, a Professora Doutora Ana Alice Baptista e o Professor Doutor Miguel Abrunhosa Brito, por todo o apoio, empenho e paciência ao longo da elaboração desta dissertação.

Desejo igualmente agradecer a todos os integrantes do projeto de investigação *IViSSEM – 6.849,32 Journal Articles Everyday: Visualize or Perish!* que me acolheram, em especial ao Professor Doutor Jorge Sá e ao Professor Doutor Bruno Azevedo. Não poderia ainda deixar de referir o apoio da Universidade do Minho, da Fundação para a Ciência e a Tecnologia e do Fundo Europeu de Desenvolvimento Regional.

Queria, da mesma forma, reconhecer a assistência do Professor Doutor Filipe Alvelos e da Professora Alexandra Gaspar.

Agradeço também aos meus colegas do Mestrado em Sistemas de Informação Camilo Moro, Diana Freitas, Eduardo Farinha e Sara Coelho. Aos meus amigos André Veiga e Tânia Ferreira, a quem prometi uma menção.

Por último, mas não menos importante, agradeço aos meus pais e irmão pelo apoio incondicional. Prapris, obrigada pelas revisões incansáveis.

Correndo o risco de injustamente não ter mencionado algum dos contributos, a todos deixo expresso o meu mais sincero e profundo agradecimento.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio, nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

Sistemas de Recomendação Científica: Incorporação da Reputação Científica nos seus Algoritmos

Inerente à investigação científica está a procura de informação ou dados que sustentem ou contribuam para os resultados de investigação. Os sistemas de recomendação científica sugerem itens relevantes aos investigadores, minimizando a sobrecarga de informação e atendendo aos seus gostos, preferências e necessidades.

A dissertação tem como objetivos (1) propor uma métrica quantificadora da reputação científica pessoal para aplicar ao sistema de recomendação científica IViSSEM, baseada numa revisão de literatura sobre a reputação científica, (2) apresentar um plano de validação para a métrica e (3) apresentar um plano de testes do sistema de recomendação científica IViSSEM, baseado numa revisão de literatura aos sistemas de recomendação científica.

As revisões de literatura seguem a diretriz *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA). A proposta da métrica e o plano de validação da mesma seguem uma metodologia híbrida do Método Delphi combinado com o *Full Consistency Method*.

A revisão de literatura aos sistemas de recomendação científica contribui para a compreensão das principais técnicas de recomendação, num contexto científico, incluindo os problemas que lhes está associado e possíveis soluções. Permite igualmente ter uma visão dos tipos de avaliação que se podem realizar aos sistemas de recomendação. O plano de testes ao IViSSEM, em especial, proporciona uma visão dos procedimentos para uma avaliação offline e estudo de utilizador. A revisão de literatura à reputação científica resultou numa proposta de definição de reputação científica pessoal. Em adição, propõe-se uma métrica de reputação científica, abrangente na sua definição e que, por esse motivo, ajuda a combater a mentalidade *publish or perish* e não exclui os investigadores com início de carreira, problemas identificados na literatura.

Palavras-chave: Avaliação dos Sistemas de Recomendação; Reputação Científica; Sistemas de Recomendação Científica.

ABSTRACT

Scientific Recommender Systems: The Incorporation of scientific reputation in their algorithms

Inherent to scientific research is the search for information or data to support or contribute to research findings. Scientific recommendation systems suggest relevant items to researchers, minimizing information overload and catering to their tastes, preferences and needs.

The dissertation aims to (1) propose a metric quantifying personal scientific reputation to apply to the IViSSEM scientific recommendation system, based on a literature review of scientific reputation, (2) present a validation plan for the metric and (3) present a testing plan for the IViSSEM scientific recommendation system, based on a literature review of scientific recommendation systems.

The literature reviews follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline. The proposed metric and its validation plan follow a hybrid methodology of the Delphi Method combined with the Full Consistency Method.

The literature review on scientific recommendation systems contributes to the understanding of the main recommendation techniques, in a scientific context, including the problems associated to them and possible solutions. It also provides an insight into the types of evaluation that can be performed on recommender systems. The IViSSEM test plan, in particular, provides insight into the procedures for an offline evaluation and user study. The literature review on scientific reputation resulted in a proposed definition of personal scientific reputation. In addition, a metric of scientific reputation is proposed, which is comprehensive in its definition and therefore helps to combat the publish or perish mentality and does not exclude early career researchers, problems identified in the literature.

Keywords: Recommender Systems Evaluation, Scientific Recommender Systems , Scientific Reputation.

ÍNDICE

LISTA DE ABREVIATURAS/SIGLAS

AHP	Analytic Hierarchy Process
AP	Average Precision
CB	Content-Based
CBF	Content-Based Filtering
CEICSH	Comissão de Ética para as Ciências Sociais e Humanas
CF	Collaborative Filtering
CGPRec	Content-Based and Knowledge Graph-Based Paper Recommendation
Cos	Cosine
CTR	Click-Through Rates
DCG	Discounted Cumulative Gain
F1	F-Measure
FDM	Fuzzy Delphi Method
FUCOM	Full Consistency Method
GB	Graph-Based
IDF	Inverse Document Frequency
JIF	Journal Impact Factor
LOD	Linked Open Data
MAE	Mean Absolute Error
MAP	Mean Average Precision
MRR	Mean Average Precision
nDCG	Normalized Discounted Cumulative Gain
ORCID	Open Researcher and Contributor ID
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
REF	Research Excellence Framework
RMSE	Root Mean Squared Error
RO	Research Output
sim	Similarity
SR	Sistema de Recomendação
TF	Term Frequency

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Este capítulo introdutório apresenta uma visão geral da presente dissertação. Na secção 1.1 é feita uma contextualização dos temas a abordar. Na secção 1.2 expõe-se a motivação, aliada ao projeto *IViSSEM - 6.849,32 Journal Articles Everyday: Visualize or Perish*!, introduzido na secção 1.3. Na secção 1.4 enunciam-se os objetivos do trabalho, complementados pelas questões de investigação. Na secção 1.5, que conclui a Introdução, ilustra-se a estrutura da dissertação, com referência breve aos capítulos que a formam.

1.1 Contextualização

A expansão da *World Wide Web* e consequente aumento de páginas web avolumou a quantidade de informação e dados digitais disponíveis online. Navegar por entre toda esta informação afeta não só a capacidade de processar informação, mas também a capacidade de avaliar a qualidade da mesma.

Num contexto académico, a era tecnológica vivenciada despoletou um aumento do número de resultados de investigação (Amami et al., 2017) produzidos e publicados sob a forma de artigos de revista, atas de conferência, blog posts, capítulos de livros, relatórios, portfólios, exibições, atuações ao vivo, gravações de podcasts e outros. Como consequência, os investigadores lidam com uma sobrecarga de informação quando procuram por trabalhos científicos relevantes e adequados à sua investigação e/ou às suas áreas de interesse (Tang et al., 2021).

Os sistemas de recomendação, em específico, os de natureza científica, ajudam os investigadores a manusear e filtrar informação com base nas suas necessidades, gostos e preferências (Ricci et al. 2015). Eles recomendam automaticamente resultados de investigação a investigadores ou outros indivíduos associados à academia com base nas informações que eles fornem ao sistema.

Diversos modelos foram já concebidos para auxiliar os utilizadores a receberem recomendações personalizadas (Beel et al. 2016). São testes aos sistemas de recomendação, contudo, que averiguam se o sistema de facto se encontra alinhado com todas as suas potencialidades.

Este trabalho sugere a incorporação da reputação científica do investigador no algoritmo de um sistema de recomendação científica para gerar recomendações e melhorar o seu desempenho.

Apesar de a reputação ser de natureza subjetiva, há uma tendência e constante preocupação para a quantificar. Todavia, a mensuração da reputação tem-se apoiado maioritariamente em citações e no

número total de publicações, conferindo uma visão estreita ao conceito de reputação. É fundamental considerar o percurso do investigador e as atividades inerentes à investigação, em especial quando as métricas tradicionais de reputação se traduzem em oportunidades e avanços na carreira.

Nesse seguimento, propõem-se a incorporação da reputação através de uma métrica quantificadora, que abrange diversos critérios, permitindo uma visão ampla do conceito. Por esse motivo, a métrica não exclui investigadores com início de carreira recente. Ademais, procura-se validá-la junto da comunidade científica, a quem ela se dirige e que, portanto, insta o seu input.

Estando-se a debruçar sobre a reputação científica, seria negligente não mencionar fatores como o país de origem do investigador e o seu estado económico, político e social; o sexo do investigador; a sua situação económica atual e passada; e o prestígio da universidade frequentada, entre outros, que criam desigualdade entre os investigadores. Eles determinam as oportunidades recebidas, o percurso de carreira, o trabalho de investigação em si e, de modo consequente, a reputação científica do investigador (Wilsdon, 2016). Os fatores enquadram-se, porém, no campo da sociologia, uma área distinta àquela em estudo, pelo que não compreendem o âmbito da presente investigação.

Salvaguarda-se ademais que nenhuma métrica é capaz de representar na totalidade o investigador e o seu trabalho (Penfield et al., 2014; Hanafy et al., 2018). Não obstante, indicadores quantitativos cuidadosamente selecionados e aplicados podem ser um complemento útil para outras formas de avaliação e tomadas de decisão (Hanafy et al., 2018).

1.2 Motivação

A presente dissertação é motivada pelo projeto IViSSEM - 6.849,32 Journal Articles Everyday: Visualize or Perish!, que visa desenvolver e testar uma nova altmetric, denominada Social Scholarly Experience Metric (IViSSEM, sem data).

A altmetric resultará da aplicação de técnicas de *machine learning* e será aplicada ao seu sistema de recomendação científica. Sistema esse que pretende utilizar a reputação do investigador no seu algoritmo e substituir as listas de resultados por técnicas vanguardistas de visualização de informação.

O algoritmo de recomendação IViSSEM visa recomendar os resultados de investigação mais relevantes aos seus utilizadores, recomendados por investigadores. Numa etapa inicial, quando o utilizador procura no sistema sobre um determinado tópico/domínio científico, o algoritmo de recomendação identifica os top 20 investigadores com mais reputação e influência nesse tópico/domínio. Analisa, de seguida, os

resultados de investigação que os 20 investigadores recomendaram e, por fim, apresenta os mais relevantes ao utilizador.

O sistema de recomendação IViSSEM é composto por 3 métricas: métrica de reputação, métrica de influência e métrica de *expertise*. A influência é composta por indicadores que espelham a credibilidade do investigador na plataforma, nomeadamente o número de procuras, número de recomendações, número de seguidores e número de pessoas que segue. Em contrapartida, a reputação é determinada por indicadores, nomeadamente prémios, v-índex, k-índex, redes de citações, patentes, emprego, financiamento, orientações (de alunos de doutoramento e mestrado), participações em eventos da comunidade académica e revisões. A métrica de *expertise* subdivide-se em *expertise* pessoal e *expertise* na plataforma. O primeiro corresponde aos níveis de conhecimento *beginner, intermediate* e *expert*; o segundo ao número de procuras e recomendações no domínio ou subdomínio.

Esta dissertação propõem uma métrica de reputação científica para integrar o sistema de recomendação, bem como um plano de avaliação ao mesmo para estudar a viabilidade da proposta.

1.3 Objetivos

A presente dissertação surge na sequência do projeto IViSSEM, abrangendo o desenvolvimento de um plano de testes, em adição à criação e validação de uma métrica de reputação para o seu sistema de recomendação. Deste modo, o trabalho tem os seguintes objetivos:

1. propor uma métrica quantificadora da reputação científica pessoal para aplicar ao sistema de recomendação científica IViSSEM, baseada numa revisão de literatura sobre a reputação científica,
2. apresentar um plano de validação para a métrica quantificadora da reputação científica pessoal e
3. apresentar um plano de testes do sistema de recomendação científica IViSSEM, baseado numa revisão de literatura aos sistemas de recomendação científica.

Para concretizar o objetivo 1, realizou-se uma revisão de literatura sobre a reputação científica pessoal. Nesse seguimento, procurou-se responder às questões de investigação:

1. O que é a reputação científica pessoal?
2. Como se mede a reputação científica pessoal?

Ambas as perguntas permitem entender a perceção que os investigadores têm de reputação científica. A pergunta 2 proporciona uma visão das medidas, métricas e critérios que estão a ser aplicados para medir o conceito. As respostas são um ponto de partida para a proposta da métrica.

Para concretizar o objetivo 2, realizou-se uma revisão de literatura sobre os sistemas de recomendação científica e os métodos de avaliação dos mesmos. A finalidade é consultar, analisar, sintetizar e reportar sobre a literatura referente aos sistemas de recomendação científica. Por conseguinte, procurou-se responder às seguintes perguntas:

1. Como funcionam as principais técnicas de recomendação científica?
2. Quais são os principais desafios das técnicas de recomendação?
3. Qual é o procedimento para a realização de uma avaliação aos sistemas de recomendação científica?
4. Que instrumentos são necessários à avaliação de um sistema de recomendação científica?

As perguntas 1 e 2 prendem-se à necessidade de compreender como funcionam os sistemas de recomendações na geração de recomendações, em especial, o sistema de recomendação IViSSEM. As perguntas 3 e 4 relacionam-se diretamente com a elaboração do plano de testes do sistema de recomendação IViSSEM.

1.4 Estrutura da dissertação

Esta dissertação desdobra-se em 6 capítulos, ordenados de forma a refletir a sequência de etapas realizada para o desenvolvimento da métrica de reputação científica e respetivo plano de validação para o IViSSEM, e do plano de testes do sistema de recomendação IViSSEM.

No capítulo 1 é feita uma introdução e contextualização do tema proposto e do projeto para o qual o trabalho foi desenvolvido; também são descritos os objetivos a alcançar.

O capítulo 2 consiste numa revisão de literatura sobre a reputação científica. Nele descrevem-se as diversas formas de medição da reputação, discute-se o papel das redes sociais científicas na reputação e a emergência de *altmetrics*. Em adição, este capítulo apresenta uma reflexão sobre os esforços e métodos para estimar a reputação, bem como uma proposta de definição do conceito reputação científica pessoal.

O capítulo 3 desvela a proposta para uma métrica de cálculo da reputação científica e o plano do processo de validação da mesma.

O capítulo 4 compreende a revisão da literatura sobre as técnicas de recomendação e tipos de avaliação dos sistemas de recomendação científica. Inclui uma apresentação dos problemas associados às técnicas, assim como as soluções encontradas por investigadores da área, e uma referência às ferramentas e instrumentos necessários à realização da avaliação do sistema.

O capítulo 5 exibe o plano de testes ao sistema de recomendação científica IViSSEM, sustentado pela revisão de literatura do capítulo 2. Descreve todo o processo de avaliação, indicando exatamente o que fazer em cada etapa e os instrumentos auxiliares.

O capítulo 6 contém as conclusões do trabalho e responde às perguntas introduzidas nos objetivos.

2. REVISÃO DE LITERATURA À REPUTAÇÃO CIENTÍFICA PESSOAL

O investigador aspira contribuir para o progresso científico, objetivo ao qual não se pode desprender a necessidade da partilha do mesmo com a comunidade e sociedade. É a comunidade científica que avalia, valida e atribui credibilidade e valorosidade aos resultados de investigação (Nicholas et al., 2018; Herman & Nicholas, 2019). Ser investigador implica, portanto, a produção de research outputs e a avaliação destes por terceiros. Por extensão, também o investigador, intrínseco ao seu trabalho, é apreciado. Desta apreciação deriva o sucesso e reputação do investigador (Nicholas et al., 2018). Pode-se dizer que a reputação é indício do desempenho e qualidade da contribuição científica do investigador. Por isso ele a cobiça.

A reputação científica e o seu cálculo são um tema de interesse crescente, em virtude das auditorias às despesas públicas no ensino superior e na investigação, das exigências para uma mensuração da qualidade e impacto da investigação, da competitividade entre instituições académicas por prestígio, e de outros fatores.

A reputação tem três as dimensões, dependendo da entidade sobre a qual a avaliação incide. O presente trabalho ocupa-se com a dimensão pessoal, ou seja, com a avaliação do investigador

A revisão de literatura surge para responder às questões de investigação: (1) o que é a reputação científica pessoal?, (2) como se mede a reputação científica pessoal?.

Nos próximos capítulos, *Metodologia da revisão de literatura à reputação científica* informa, como o título indica, da metodologia na qual esta revisão literária se baseia; A *reputação científica* introduz a definição do conceito em estudo, alude às referidas dimensões da reputação, e refere a necessidade e importância de medir a reputação; *Medir a reputação científica* expõe as diferentes maneiras de calcular a reputação do investigador, nomeadamente através de *peer reviewing*, nas redes sociais científicas e com as métricas tradicionais e alternativas; *Comentários sobre a medição da reputação* é o produto de uma reflexão sobre a apreciação da reputação; e *Proposta de uma definição para a reputação científica pessoal* surge como um resultado da revisão de literatura.

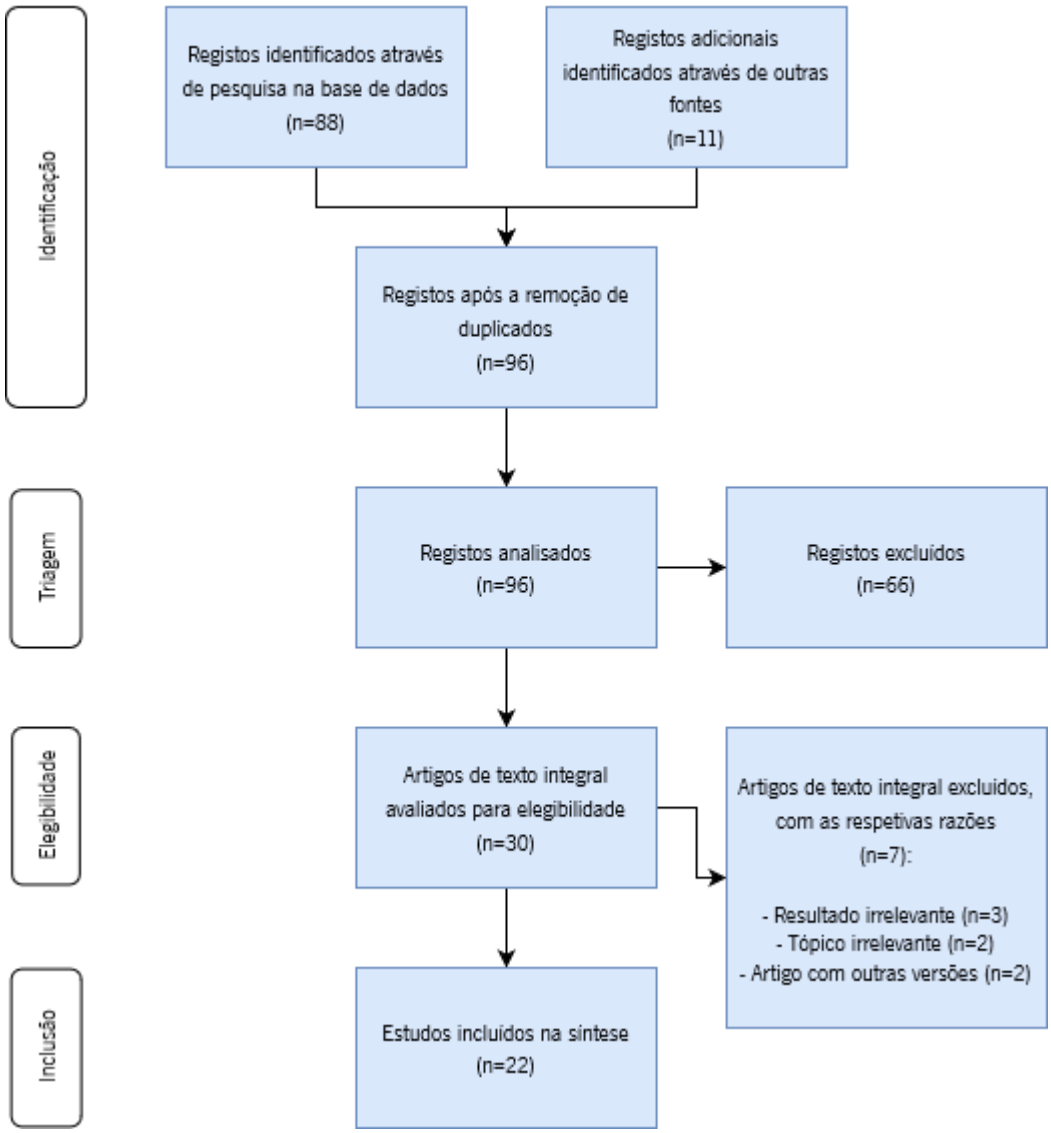
2.1 Metodologia da revisão de literatura à reputação científica

Com o objetivo de criar uma métrica de reputação, nasce a necessidade de responder à questão de investigação: "quais são os critérios que determinam a reputação científica?" Pretende-se, assim, indagar

a natureza atual da reputação científica e aferir critérios úteis para a definição da métrica de reputação a propor.

O processo adotado para encontrar resposta foi a revisão de literatura, seguindo a diretriz PRISMA de Moher et al. (2009). A figura 9 contempla o diagrama de fluxo PRISMA aplicado ao trabalho de investigação sobre a reputação científica, e é através dela que se se passa a explicar a metodologia da revisão de literatura.

Figura 1 Fluxograma da revisão de literatura à reputação científica.



Fonte: Adaptado de Moher et al. (2009)

Sucintamente, num primeiro momento, foi feita uma seleção preliminar de artigos científicos com a adequação do título e resumo enquanto critérios de elegibilidade/exclusão, auxiliada por uma base de dados. Procedeu-se com uma leitura integral dos documentos, a qual determinou os registos incluídos na revisão de literatura.

A identificação das publicações sobre o tópico em causa foi feita com recurso à base de dados Scopus, uma das três bases de dados bibliográficas multidisciplinares mais importantes. Deu-se preferência ao Scopus por possuir uma cobertura da literatura científica mais extensa do que a Web of Science (Wilsdon, 2016) e oferecer melhor usabilidade e opções de configuração da pesquisa e dos resultados decorrentes do que o Google Scholar.

Os termos de pesquisa aplicados foram as alternativas scholarly, researcher e academic, bem como a keyword reputation. Foram também impostas as condições de os resultados estarem em inglês e terem sido publicado após 2012. A pesquisa efetuada pode ser reproduzida utilizando [2] como guia:

> [1] TITLE (academia OR researcher OR scholarly) AND TITLE (reputation) AND LIMIT-TO (LANGUAGE, "English") AND LIMIT-TO (PUB YEAR > 2012)

Da pesquisa derivou uma lista de 88 registos, à qual se acresceu outros 10 por efeito do processo de *backward reference searching* (9), procura de um artigo específico da autoria do inventor do h-índex (1) e recomendação pessoal (1). Como último passo da fase identificação, excluiu-se 3 registos duplicados. Na fase de triagem, dos 96 registos analisados, 66 foram removidos devido à sua irrelevância, comprovada pelo título e/ou conteúdo do resumo. Assim sendo, no total, 30 artigos foram analisados na sua íntegra, dos quais 23 foram incluídos na revisão de literatura. Os restantes foram extraídos por (1) os resultados ou conclusões da investigação não serem relevantes para o trabalho em questão, (2) o tópico da investigação não ser relevante, e (3) os artigos serem versões expandidas de artigos publicados posteriormente pelo mesmo autor.

2.2 Definição de reputação científica na literatura

A reputação científica é para Herman (2018) a avaliação agregada que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade relativamente à produtividade e ao impacto dos seus resultados de investigação.

A produtividade é, em parte, a conciliação dos conceitos quantidade e qualidade, aplicados aos resultados de investigação. Quanto maior e melhor as componentes, maior a produtividade; quanto maior a produtividade, mais favorável a reputação.

O impacto é o poder transformativo do resultado de investigação na ciência, na sociedade (Penfield et al., 2014), na economia, na cultura, na saúde, nas políticas, nos serviços públicos, no ambiente. É o efeito dos *research outputs* na comunidade científica e/ou na comunidade científica e/ou na comunidade em geral.

Na definição dada, a reputação científica é aplicada a entidades singulares. Na verdade, são três as dimensões da reputação, consoante a entidade sobre a qual a avaliação incide: (1) dimensão pessoal, enquadrada no trabalho individual; (2) organizacional, enquadrada no trabalho em unidades de Investigação e Desenvolvimento; (3) interorganizacional, enquadrada no trabalho em parcerias empresariais (Ruão & Pessôa, 2019). A presente secção preocupa-se com a dimensão pessoal.

Numa outra perspetiva, a reputação simboliza a ideia que as várias comunidades têm sobre os investigadores e os seus ROs. Essa ideia resulta de interações (Nicholas, 2017) unilaterais (e.g.: apresentações em conferência) e bilaterais (e.g.: reuniões de um projeto de investigação). As interações tanto podem ser vividas pela própria pessoa ou por um terceiro, que as relata.

A própria definição de reputação remete para a natureza subjetiva do fenómeno (Cleary et al., 2012). Uma avaliação, ainda que guiada por critérios, é sempre subjetiva.

Estando-se a viver numa sociedade em rede, os investigadores, que se conectam com os seus semelhantes e com organizações numa escala global, vêm a reputação influenciada por essas relações. Ciclicamente, a reputação influencia a rede de contactos. Devido à dinâmica do meio, o investigador nunca está isolado. Também das relações estabelecidas entre investigador-investigador, investigador-organização e investigador-ambiente surgem novos resultados de investigação.

2.3 Medir a reputação científica

A reputação tem aspetos quantitativos e qualitativos. Para avaliar estes últimos, um método que envolva subjetividade, como a avaliação por pares, parece indicado. Contudo, esta abordagem é recursos-intensiva e, por isso, nem sempre aplicável.

Nesse contexto, as métricas e indicadores quantitativos são instrumentos que simplificam o processo. Se, por um lado, a sua utilidade é indiscutível, por outro, a sua simplificação (do conceito de reputação) é a sua principal desvantagem. A reputação inclui subjetividade, difícil de captar em técnicas bibliométricas.

Numa altura em que é habitual julgar a reputação quantitativamente (Cleary et al., 2012; Wilsdon, 2016), crescente é a importância de entender de que forma ela pode ser medida, e quais as alternativas. Não só está em causa o investigador e o seu trabalho mas a sua progressão na carreira, a aprovação dos pares, a obtenção de financiamento, a conexão com colegas, o apoio de agências governamentais. Atualmente, as métricas tradicionais de reputação decidem a quem é oferecido uma posição, quem é

promovido no trabalho, quem recebe recursos, quem é gratificado com um prémio, quem tem direito a uma bolsa de investigação (Herman & Nicholas, 2019).

A preocupação em quantificar o investigador é exacerbada por auditorias às despesas públicas no ensino superior e na investigação, pelo quesito de mensurações mais estratégica da qualidade e impacto da investigação, pela competitividade entre instituições por prestígio, entre outros (Wilsdon, 2016). O propósito destas exigências é compreensível: o financiamento é finito e é daí emerge a necessidade de otimizar a alocação de recursos.

Paralelo à preocupação constante em quantificar a reputação deveria estar a preocupação em entender o que medir e como medir a reputação.

2.3.1 Peer reviewing

Medir o impacto, considerando a sua definição, teria de envolver uma contagem do número de pessoas singulares e coletivas cujo pensamento e/ou prática foi alterada por um RO. Ainda que se possa reconhecer o impacto de um indivíduo na ciência, na economia, etc., é tarefa árdua escolher os indicadores adequados. Nenhum é desprendido de limitações ou desvantagens, incluindo a sua aplicação impertinente em diferentes situações e incapacidade de expressar todo um contexto.

Uma forma de captar a mudança de opinião, comportamento e práticas como resultado da investigação é através da recolha de testemunhos ou da realização de inquéritos/entrevistas (Barker et al., 2011; Penfield et al., 2014). Barker et al. (2011) identificaram entrevistas a *stakeholders* como sendo um método viável para avaliar o impacto social de investigação, tendo desenvolvido um questionário auxiliar. De forma complementar, delinearam uma abordagem de 4 etapas para tornar a relevância social passiva de avaliação para investigadores e *peer committees* em auditorias universitárias.

O método acima descrito é uma forma de *peer reviewing*, um dos métodos mais importantes na avaliação (Hirsch, 2005; Nicholas, 2017; Ruão & Pessôa, 2019) por satisfazer a parte subjetiva da reputação e combinar métodos qualitativos.

Geralmente, um processo que envolve painéis de especialistas (ex: *peer reviewing*) vale-se de métodos quantitativos e, frequentemente, qualitativos, como é o caso das entrevistas. Ou seja, viabiliza uma abordagem híbrida.

No Reino Unido, o *Research Excellence Framework* (REF) é utilizado para avaliar a qualidade da investigação em Instituições do Ensino Superior, sendo que os resultados determinam o financiamento futuro de cada instituto. Apesar de não pertencer à dimensão pessoal da reputação, trata-se de um

processo de avaliação de pares, auxiliado por indicadores quantitativos, no qual investigadores são escrutinados.

O processo de avaliação do REF consiste na formação de painéis de especialistas, constituídos por académicos seniores, membros internacionais e utilizadores de investigação. Para cada submissão, são avaliados três critérios: (1) a qualidade dos outputs; (2) o impacto na economia, sociedade e/ou cultura; e (3) o ambiente da investigação (England, sem data; REF 2014, 2014).

Os outputs podem ser do tipo: livros (ou partes de livros, como capítulos e notas introdutórias), artigos de revista e contribuições de conferência, artefactos físicos, exibições e performances, artefactos digitais (incluindo conteúdo web), e outros. Revisões, manuais ou obras editadas podem ser incluídas, mas teses, dissertações e artigos submetidos para conclusão de um grau académico não se incluem nos outputs aceites (REF 2014, 2014).

Na verdade, os outputs ou resultados de investigação são imperativos para a reputação visto o objetivo ou um dos objetivos primários do investigador ser a criação de novo conhecimento (Jamali et al., 2016; Ruão & Pessôa, 2019). Aos outputs são associados indicadores, tendo como exemplo (REF 2021, 2021):

- Citações num debate público,
- Citações em documentos de política, regulamentação, estratégia, etc.,
- Indicadores quantitativos ou estatísticas sobre o número de participantes num evento de investigação,
- Feedback qualitativo dos participantes em eventos de investigação,
- Agradecimentos aos investigadores em páginas web, relatórios ou briefings,
- Citações diretas da investigação em publicações parlamentares, tais como relatórios de comissões, ou briefings,
- Dados que demostrem relações de trabalho próximas como o número de reuniões realizadas e coautores de publicações,
- Depoimentos de membros, comissões ou funcionários,
- Dados para mostrar relações de trabalho próximas (o número de reuniões realizadas e coautores de publicações, por exemplo),
- Testemunhos de membros, comissões ou funcionários,
- Provas de utilização de processos/tecnologias resultantes da investigação,
- Acordos formais de parceria ou colaboração em investigação com instituições, ONGs e organismos públicos,

- Trabalho de consultadoria,
- Citações por jornalistas, emissoras ou meios de comunicação social,
- Debate público nos meios de comunicação social,
- Uso de materiais educativos resultantes da investigação.

Para um compêndio de indicadores exemplificativos a fonte REF 2021 (2021) é conveniente. São aqui reproduzidos uma pequena porção, que convém a natureza híbrida do *framework* e aplicabilidade no domínio da ciência da computação.

Como é passível de se observar, as citações – uma métrica tradicional (Herman & Nicholas, 2019; Nicholas, 2017) – são consideradas, em conjunto com outros critérios, numa iniciativa de alargar o entendimento convencional de reputação e mitigar o problema que surge quando apenas uma quantidade limitada de critérios é ponderada (Jamali et al., 2016; Nicholas, 2017).

No contexto de outputs, para além de indicadores baseados em publicações como livros e capítulos de livros (escritos, editados, traduzidos), do número de participantes num evento, do número de visitantes a uma exibição e do uso de materiais de educação,

- a posição do nome de determinado investigador na lista de autores de uma publicação,
- as coautorias,
- o número de postos de trabalhos oferecidos aceites,
- indicadores de avaliação do impacto de citação dos jornais (nomeadamente o *Journal Impact Factor* (JIF) e *Eigenfactor*) e
- o lucro gerado com a venda de ROs ou produtos/serviços derivados do trabalho de investigação

são todos indicadores de reputação (Penfield et al., 2014; REF 2021, 2021; Wilsdon, 2016).

Na vertente de *peer reviewing*, é de mencionar as plataformas dedicadas à atividade ou que permitem, ao investigador, evidenciar a sua contribuição para a atividade, e aos seus colegas, validar este seu exercício. Projetos como o Publons[1] motivam a criação de um perfil, sujeito a análise por um empregador ou organização.

[1] http://publons.com/

2.3.2 A emergência de redes sociais académicas

Neste momento, medir qualitativamente a reputação - que se traduz em *peer reviewing* - é mais complexo e intensivo, em termos de recursos. Por isso o seu uso em avaliações pontuais. Em plataformas/base de dados científicas ou académicas, onde indicadores e métricas são intencionalmente consultáveis a qualquer momento e recalculados ou atualizados automaticamente, índices bibliométricos concentrados em citações e publicações são das ferramentas mais utilizadas para avaliar o desempenho e contribuição do investigador, e medir a qualidade da investigação (Wilsdon, 2016). Devido, talvez, em parte, à facilidade da obtenção dos valores em bases de dados. Plataformas à semelhança do Publons surgem como solução para a dicotomia qualidade-quantidade.

De facto, a emergência de redes sociais científicas dá azo a novas formas de medição da reputação. Redes sociais científicas conectam pessoas – o que pode resultar em colaborações – e permitem ao utilizador evidenciar o seu desempenho científico, convertendo interações e ROs em reputação (Herman & Nicholas, 2019).

Tome-se o exemplo do ResearchGate que, além de indicadores tradicionais, considera ainda atividades específicas à plataforma: *questions and answers* (Q&As), número de seguidores, número de visualizações do perfil e endossos das competências/conhecimentos pelos pares (Nicholas et al., 2016). Retratam elas a influência que um investigador tem na plataforma. Significa isto que os aspetos mais sociais do ResearchGate não representam verdadeiramente a reputação do investigador no e fora do meio académico/científico, mas a reputação na rede social. É, por isso, relevante estabelecer essa diferença. Se a reputação de um investigador numa plataforma científica retrata a atividade científica interna e externa à plataforma, é justificado questionar a utilidade do ResearchGate e afins na carreira dos investigadores. A atividade interna atividade pode introduzir o trabalho de investigador a novas pessoas, formar novas relações, ocasionar colaborações e distender a rede de citações, tendo impacto na vida real. Conquanto, a atividade externa, composta por indicadores quantitativos, pode ser consultada noutras fontes secundárias, das quais já se referiu a Web of Science e o Scopus.

Na prática, as opiniões dos investigadores e a perceção das instituições/empregadores diferem em relação à reputação. Os empregadores não valorizam a divulgação da investigação através de blogues e redes sociais de igual modo aos investigadores (Jamali et al., 2016). Inclusive, alguns investigadores duvidam das métricas próprias das redes sociais/plataformas emergentes e a sua contribuição para a reputação (Jamali et al., 2016).

2.3.3 Métricas tradicionais da reputação científica

As métricas tradicionais baseiam-se maioritariamente no número total de publicações e no número total ou média de citações. Ao nível do autor e ao nível de artigos (Herman & Nicholas, 2019; Nicholas, 2017; Wilsdon, 2016). Segundo um estudo de Aung et al. (2019), as métricas tradicionais mais familiares e utilizadas são o JIF, o número total de citações e o h-índex.

O JIF de uma revista científica é o número médio de citações por artigo publicado num período de 2 anos. O h-índex é uma proposta para quantificar a produtividade e o impacto dos investigadores, baseada nas citações e publicações. Corresponde ao número de publicações (h) de um investigador com, pelo menos, o mesmo número (h) de citações (Hirsch, 2005). Ou seja, um investigador tem h-índex = 16 quando 16 das suas publicações têm, pelo menos, 16 citações.

O uso extenso das mencionadas métricas tradicionais não significa que estejam isentas de críticas negativas (Hanafy et al., 2018; Hirsch & Buela-Casal, 2014).

Sabendo que o número de publicações é significativo, os investigadores podem publicar os seus ROs em partes para obter um maior número de contagem, publicar colaborativamente com um grande número de colegas, diminuindo a sua carga de trabalho no projeto. Nomes reconhecidos são acrescentados à lista de autores com o objetivo de aumentar a visibilidade e as probabilidades de publicação, mesmo que estes em nada ou pouco tenham contribuído (Fong & Wilhite, 2017).

Um foco copioso nas citações motiva o ato de citações próprias. Noutras instâncias, os investigadores são coagidos por editores a acrescentar citações dos artigos de revista deles (Fong & Wilhite, 2017). Se os valores destas duas medidas estão comprometidos, também os valores do JIF e do h-índex estão. O h-índex sofre ainda de outros problemas. É dependente da área da ciência e da popularidade de um tema ou campo de investigação, e é inadequando em situações em que o investigador tem um número baixo de publicações mas com elevado número de citações (Hanafy et al., 2018; Hirsch & Buela-Casal, 2014).

Hanafy et al. (2018) questionam a validade das métricas tradicionais ao reconhecer a variação no número de citações entre os diversos domínios de investigação; as disparidades entre o número de citações do mesmo investigador nas ferramentas bibliométricas Scopus, Web of Science e Google Scholar; e o impacto da rede de contactos nas citações.

Não obstante, a literatura mostra que o abuso ou apoio excessivo das citações em momentos de avaliação da reputação não nega o impacto dos resultados de investigação no desenvolvimento de novos processos, ideias, bens, métodos, legislação e políticas (Penfield et al., 2014; REF 2021, 2021). Ainda

para mais, quando as citações são evidenciadas por citações em documentação externa ao ambiente académico, como atas das sessões e debates parlamentares, atas de reunião de comissões, atas da assembleia da república, e relatórios jurídicos.

Citações nos meios de comunicação social como redes sociais, websites, microblogs são hoje em dia tidas em conta, em grande parte, devido à popularização do conceito das *altmetrics* (ou métricas alternativas).

2.3.4 Métricas alternativas da reputação científica

Tendo o mundo digital importância crescente na vida em geral mas, sobretudo, na ciência, desponta-se a necessidade de incluir em avaliações outputs e dados advindos de outputs específicos ao meio. Designadamente, o número de downloads de outputs, o número de visitantes num blog, o número de leitores de um blog, o tempo passado a ler um output, e o número de ouvintes de um output (p.e., podcasts) (Wilsdon, 2016).

As *altmetrics* surgem como complemento às métricas tradicionais (tabela 1). Elas medem as interações académicas que sucedem na Web em canais de comunicação como redes sociais, blogs e gestores de referências bibliográficas. Visam, portanto, medir o nível de discussão e partilha *research outputs*. Isto é, a visibilidade no meio online.

Elas permitem remover o foco inflexível até há pouco dado às métricas tradicionais. Graças a elas, os investigadores têm agora acesso a uma maior quantidade e diversidade de métricas, que permitem uma melhor compreensão da forma como a investigação está a ser consumida, debatida e disseminada.

Tabela 1 Exemplos de métricas tradicionais e de métricas alternativas.

Métricas tradicionais	**Métricas alternativas**
Número total de citações (com e sem autocitações)	Número de visualizações e downloads do RO
Número total de ROs	Número de seguidores
Journal Impact Factor (JIF)	Número de reviews
H-índex	Número de "gostos" nas redes sociais
Eigenfactor Score	Número de partilhas nas redes sociais
Journal Cited Half-Life	Número de menções na Wikipédia
SCImago Journal Rank	Número de bookmarks

Fonte: (Aung et al., 2019)

São empresas como a Altmetrics[2] e a PlumX[3], que se dedicam a medir a influência da investigação científica, e ferramentas como a ImpactStory[4], onde o investigador pode explorar e partilhar o impacto online da sua investigação, que possibilitam o tal acesso, bem como a propagação da visão de uma reputação mais compreensiva.

Apesar de uma discussão crescente sobre as *altmetrics* e a sua relação com as métricas tradicionais, as métricas tradicionais são mais conhecidas e, por isso, empregues pelo investigador. Entre as possíveis explicações encontram-se a novidade do conceito, a falta de benefícios apercebidos do uso de *altmetrics*, a ausência de recompensa ou encorajamento de instituições na avaliação de performance e promoções, e a preocupação de invasão de privacidade (Aung et al., 2019).

É ainda de notar que, como aludido anteriormente em A emergência de redes sociais científicas, os céticos não consideram as atividades nas redes sociais relevantes para o trabalho científico, enquanto outros consideram que são, sim, relevantes visto as redes sociais ajudarem a alcançar um público mais vasto (Jamali et al., 2016).

O estudo de Ruão & Pessôa (2019), revela que, para os investigadores, é pertinente comunicar o trabalho individual junto de colegas e do o público em geral. Em simultâneo, os mesmos somente se preocupam com a divulgação da investigação nos meios académicos, ignorando as redes sociais e outros meios de comunicação frequentados pelo público em geral. Um questiona-se se este facto e a convicção da irrelevância das redes sociais (científicas e outras) não estão relacionados com a postura reservada dos institutos/empregadores para com o uso profissional de meios de comunicação académicos ou dirigidos ao público em geral.

Em acréscimo aos indicadores já mencionados, outros podem ser ponderados, desde que eles respeitem a definição de reputação e incidam sobre os construtos que a envolve. Qualquer aspeto vital do processo da investigação tem potencial para se transformar em critério da reputação científica pessoal. Por isso também se observam diferentes iniciativas com critérios variáveis. É, todavia, preciso um consenso entre a organização responsável, os seus membros, os participantes voluntários e os ademais envolvidos, relativamente à perspetiva promovida pela iniciativa.

[2] https://altmetric.com/

[3] https://plumanalytics.com/

[4] https://profiles.impactstory.org/

Nessa medida, o financiamento (por intermédio de uma entidade pública ou por uma indústria) e as bolsas de investigação apossadas são uma medida do sucesso da performance e, portanto, da reputação (Herman & Nicholas, 2019).

Na ótica do investigador - independentemente da sua idade, género ou área da ciência, a realização de investigação -, as atividades da investigação em si, a divulgação dos resultados da investigação através de artigos de revistas e livros, seguido de conferências, são as atividades de investigação que mais contribuem para a reputação.

Visto os investigadores envolverem-se hoje em atividades comerciais da ciência (Ruão & Pessôa, 2019), os resultados decorrentes - que podem tomar a forma de patentes, por exemplo - devem igualmente ser alvo de apreciação crítica.

2.4 Comentários sobre a medição da reputação

Indicadores não tradicionais medem atividades fora da investigação, que tem sido um dos apelos provindos de investigadores (Nicholas, 2017), juntamente com a mitigação do uso de indicadores tradicionais enquanto marcas de reputação. Não no sentido de extinguir o seu uso mas de diminuir a importância dada ao número de publicações e de citações, muito proeminente com a mentalidade de *publish-or-perish*. Querem-se avaliações que contemplem a trajetória (profissional) do investigador, fatores específicos ao investigador, atividades que não só a investigação e a contagem de uma maior variedade de outputs (Cleary et al., 2012; Nicholas, 2017; Nicholas et al., 2018).

Contudo, nenhum conjunto de indicadores é a solução única pois nenhum reflete verdadeiramente a reputação do investigador. Poder-se-á aproximar-se da solução ideal quando ela se assemelhar à visão dos investigadores supracitada, que sofrerá igualmente de limitações.

Uma delas é a falta de um consenso universal quanto à definição de reputação científica – não é de confundir uma definição clara e objetiva com opiniões apresentadas ao longo de um ou mais ROs. Aquilo que um considera prova de reputação, aquilo que um diz medir a produtividade e o impacto da investigação/investigador pode não espelhar a opinião de outro. Desse mesmo modo, pode haver um simples desacordo de pareceres quanto ao que é um indicador indicado (ou não) válido da reputação. É, pelo menos, necessário um consenso entre as partes envolvidas quando está em causa uma avaliação enquanto não existir uma definição reconhecida e aceite pela comunidade científica.

Também alguns indicadores são facilmente manipuláveis. Como visto em *Métricas tradicionais da reputação científica* e *Métricas alternativas da reputação científica*, o número de publicações, as citações

e as métricas dependentes vinculadas às citações (p.e. h-índex e JIF) são problemáticos. Por serem frequentemente utilizados por investigadores, institutos, agências, entre outros, é plausível que estimule uma pressão que leva a uma conduta antiética.

Relacionado com as métricas alternativas, não se podia deixar de mencionar a objeção que alguns investigadores têm no que tange as métricas das redes sociais nem a relutância de instituições/empregadores/investigadores em dar-lhes mais atenção e destaque. É preciso ponderar se esta atitude se deve à relativa novidade do conceito ou se há algo mais. Uma hipótese é o sentimento dos investigadores ser negativo por os fazer sentir que estes indicadores são mais uma atividade que terão de desenvolver, e que lhes remove tempo das atividades que julgam ser mais produtivas e importantes para a investigação. Nesse seguimento, quão legítimo seria impor aos investigadores uma presença online profissional. Por outro lado, os investigadores noviços, que estão em desvantagem, podem recorrer às redes sociais para se afirmarem na comunidade (Herman, 2018).

Um outro aspeto crucial na identificação de critérios de reputação científica a serem aplicados em métricas é o acesso aos dados necessários.

A Web of Science, o Scopus e o Google Scholar são as três bases de dados bibliográficos multidisciplinares mais notáveis (Wilsdon, 2016). Devido à sua relevância são estas as ferramentas mais utilizadas pelos investigadores para tornar público os seus ROs, ver que documentos e que autores os citam, entre outros.

Numa métrica, os valores dos indicadores que a compõem são extraídos a partir de uma ou mais fontes de dados. Logo, a métrica fica restrita não só pelo acesso (ou falta de) à API das fontes, mas pela qualidade dos indicadores. Comparando todas as ferramentas anteriormente mencionadas, notam-se discrepâncias a nível de contagem de ROs, de citações, número de documentos citadores, h-índex, etc. para o mesmo investigador (Hanafy et al., 2018).

Tal ocorre porque, na maioria dos casos, a identificação da autoria das publicações e outros é automático e, como tal, suscetíveis a erros. A menos que os investigadores façam um trabalho manual e/ou monitorizem as bases de dados e plataformas de gestão do currículo para se certificarem da validade dos dados, nada garante a legitimidade total destes.

Com a emergência de plataformas de gestão do currículo, das quais se destaca o *Open Researcher and Contributor ID* (ORCID), global, e o Ciência Vitae, nacional, uma maior diversidade de dados encontra-se disponível.

Um currículo manifestamente pinta um retrato mais fiel do percurso do investigador. A informação é mais completa e dados são inseridos manualmente pela pessoa (também a possibilidade de extrair

alguns de bases de dados, como o Scopus, conectadas à conta). Os empregos, as qualificações, educação, *memberships* etc estão facilmente acessíveis.

Um dos potenciais deste tipo de plataformas é que atividades fora da investigação podem ser pesadas na avaliação da reputação, um dos embargos da comunidade.

2.5 Proposta de uma definição para a reputação científica pessoal

Não existe uma definição universalmente aceite de reputação científica na comunidade. A visão de Herman (2018) surge de outros investigadores. Corresponde, de modo igual, à opinião de muitas das referências literárias mencionadas nesta dissertação. Contudo, da revisão de literatura à reputação, conclui-se que a definição apresentada inicialmente (consultar A reputação científica) não é abrangente o suficiente para abarcar as diversas perspetivas encontradas na literatura.

Ao longo da revisão de literatura foram apresentadas ideias de autores que não se enquadram, pelo menos na totalidade, na definição dada. Como tal, propõem-se nesta secção uma definição de reputação científica pessoal baseada na revisão de literatura. É uma expansão da definição concedida por Herman (2018) e influenciada, em especial, pelos trabalhos de Cleary et al. (2012), Herman & Nicholas (2019), Nicholas (2017), Ruão & Pessôa (2019) e Wilsdon (2016).

Para efeitos da presente investigação, considera-se a reputação científica pessoal:

1. a avaliação agregada que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade em geral relativamente à produtividade e ao impacto dos seus resultados de investigação.
2. o resultado das avaliações que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade em geral relativamente à produtividade e impacto dos seus resultados de investigação, e percurso de vida profissional.
3. o produto de todo o percurso profissional transposto pelo investigador científico e que define as suas práticas, metodologias e pensamento enquanto impulsionador da ciência.

O ponto 1 já é familiar, os pontos 2 e 3 são a extensão proposta para a definição de reputação científica. Embora o ponto 1 seja, numa perspetiva pessoal, veracidade, não confere espaço para contemplar os resultados da avaliação do investigador como indícios da reputação. Por exemplo, uma agência financiadora pode avaliar um investigador, os seus ROs e, partindo daí, oferecer ou rejeitar financiamento.

A decisão de financiar, ou não, o investigador é indicação da reputação dele na perspetiva da agência. Logo, também os resultados das avaliações que o investigador recebe devem ser a sua reputação. Outro exemplo é quando uma organização oferece uma posição a um investigador, essa decisão decorre de uma avaliação. Contudo, essa avaliação não recai necessariamente apenas sobre os resultados de investigação publicados, mas da sua trajetória profissional como empregos prévios e qualificações.

Nessa veia, o produto de todo o percurso profissional do investigador – que define as práticas, metodologias, pensamento de investigação desenvolvido ao longo da carreira – é a sua reputação. Por produto insinua-se serviços prestados a outrem (p.e., voluntariado), a propriedade intelectual registada, a publicação de ROs etc. Distingue-se do resultado das avaliações porque é externo à avaliação de terceiros.

Para efeitos da presente investigação, daqui em diante, quando o termo reputação científica (pessoal) é mencionado, está-se a fazer alusão à definição de reputação científica pessoal proposta nesta secção, ou seja, aos 3 aspetos aqui explicitados.

3. PROPOSTA DE UMA MÉTRICA DE REPUTAÇÃO CIENTÍFICA

A presente dissertação surge na sequência do projeto *IViSSEM - 6.849,32 Journal Articles Everyday: Visualize or Perish*, que nasceu da necessidade sentida pelos investigadores em distinguir research outputs de interesse e qualidade por entre uma miríade de artigos científicos (Johnson et al. 2018).

Devido à abundância de informação e aumento constante de trabalhos de investigação disponíveis na web, o processo de procura e seleção de ROs relevantes aos investigadores fica dificultado. Uma das abordagens para combater este problema é o uso de sistemas de recomendação.

Esta secção do trabalho foca-se na reputação dos investigadores que, na plataforma IViSSEM, é calculada através de indicadores quantitativos respeitantes à formação e percurso profissional do investigador, produtividade, projetos e reconhecimento do trabalho. Não se centra somente em bibliometria, mas também na trajetória de vida profissional dos investigadores ao longo do tempo.

Em contraposição, a influência é composta por indicadores não tradicionais relacionados com a reputação do investigador na plataforma, nomeadamente número de seguidores, número de partilhas, número de procuras, número de pessoas que segue. Faz-se esta distinção para se entender o porquê de a reputação não incluir este género de *altmetrics*. O projeto separa o investigador profissionalmente social e influenciador do investigador profissionalmente criador (de conhecimento, sistemas, etc.).

Com fundamento na investigação e revisão de literatura realizada sobre a reputação científica, propõe-se (1) uma métrica quantificadora da reputação científica pessoal para as áreas de Ciências da Computação, Informática, Sistemas de Informação e afins, e (2) um plano de validação da métrica junto da comunidade científica.

A mensuração da reputação científica tem-se apoiado maioritariamente em métricas tradicionais que conferem uma visão estreita ao conceito. Devido ao seu amplo uso, elas não podem ser completamente ignoradas. Não podem, no entanto, ser utilizadas exclusivamente no cálculo da reputação do investigador. É fundamental considerar o percurso do investigador e as atividades inerentes à investigação. Torna-se ainda mais urgente quando estas métricas se traduzem em ofertas de emprego, promoções de trabalho, cedência de recursos, atribuição de prémios e atribuição de bolsas de investigação (Herman & Nicholas, 2019).

Propõe-se no presente trabalho uma nova métrica de reputação científica, abrangente na definição de reputação. A proposta leva em consideração o percurso profissional do investigador e não exclui aqueles com início de carreira recente. Sendo composta por multicritérios, ela cede espaço a diferentes vertentes

da reputação, valorizando determinadas particularidades que a comunidade científica tem vindo a solicitar.

Frisa-se, no entanto, que nenhuma métrica é capaz de representar na totalidade a reputação do investigador (Cleary et al., 2012; Penfield et al., 2014; Hanafy et al., 2018). Elas podem, porém, ser um complemento útil para outras formas de avaliação e tomadas de decisão (Hanafy et al., 2018).

Nas páginas que se seguem, a explicação da métrica de reputação científica pessoal proposta e o plano para a obtenção e validação da mesma é delineado. O capítulo *Trabalho prévio* referencia o trabalho realizado por alguns autores na área da reputação científica; *Apresentação da métrica de reputação* introduz e explica a proposta da métrica de reputação científica para o sistema de recomendação IViSSEM; *Definição dos critérios de reputação* apresenta os critérios englobados na métrica proposta e a razão para tal; *Definição da população de investigadores do conjunto de dados* menciona o processo de obtenção de um conjunto de dados composto por investigadores, a ser utilizado na validação da métrica; *Processo para a definição dos pesos dos critérios de reputação* descreve e esclarece o plano para a obtenção e validação da métrica de reputação proposta.

3.1 Trabalho prévio

Kalachikhin (2019) propôs um modelo de classificação da reputação de investigadores, que considera componentes positivas e negativas da reputação, sendo que o resultado (a reputação) pode ter valor nulo, positivo ou negativo. As componentes positivas são a atividade de publicação do autor, palestras em conferência, menções das publicações do investigador para além das citações, aumento do status, prémios por feitos científicos. As componentes negativas são as publicações retiradas, referências a publicações retiradas, diminuição do status, violação da ética científica e emigração. As escalas para os indicadores das componentes são determinadas a partir de métodos de *peer review*.

Hanafy et al. (2018) propuseram um modelo de confiança para conferir a produtividade do investigador, composto parcialmente por uma pontuação de autor investigador. A equação que a representa é constituída pelo número total de citações, número total de publicações, h-índex e coautorias.

Cervi et al. (2013) propuseram uma métrica multicritério que engloba bibliométricas tradicionais e o percurso de investigador enquanto profissional. A reputação é definida pela soma de 5 categorias - identificação, consultoria, júri de exame, adesão a associações, produção. Cada categoria tem diferentes elementos. Por exemplo, a produção inclui artigos em revistas científicas, capítulos de livros publicados, livros publicados, trabalhos completos publicados em conferência, h-índex, rede de coautorias, projetos

de investigação e software. O resultado de cada categoria é obtido pela multiplicação do valor do elemento pelo peso do elemento – definido pelo autor –, dividido pelo valor máximo do elemento. Assim é calculada a reputação.

Kardam et al. (2016) propuseram o ranking do trabalho académico com base na reputação do autor. Calculada através de citações, ao autor é atribuído um peso igual ao número de citações recebidas pela sua publicação. Este número pode ser dividido entre os autores dos artigos. Na eventualidade de o RO ser coautorado, o autor tem o direito de lhe ser destinado um peso próprio, decidido pelo indivíduo que o cita.

Osman & Sierra (2016) propuseram um modelo de reputação científica baseada em *peer reviews* para avaliar a reputação de investigadores e respetivos trabalhos de investigação. A reputação deriva da opinião de membros da comunidade científica sobre os artigos do investigador. Os investigadores são, então, reputados enquanto autores e *reviewers*.

3.2 Apresentação da métrica de reputação científica

A métrica de reputação proposta calcula a reputação do investigador científico com base em 20 critérios. A métrica abrange indicadores bibliométricos, mas também aspetos essenciais à carreira de investigador e à realização de investigação.

Ela contempla o percurso profissional do investigador e não exclui aqueles com início de carreira recente, tornando-a mais abrangente na definição de reputação em comparação ao trabalho prévio de Hanafy et al. (2018) e Kardam et al. (2016). Toma também uma abordagem distinta às de Kalachikhin (2019) e Osman & Sierra (2016), na medida em que a aplicação das métricas têm propósitos diferentes. A métrica proposta não depende de uma avaliação qualitativa visto o objetivo ser implementá-la numa plataforma científica digital onde a reputação é calculada automaticamente pelo sistema e não em momentos pontuais. A conceção de Cervi et al. (2013) é a que mais se assemelha à métrica proposta, com a diferença de que ela não é recebe input nem é validada pela comunidade científica.

A reputação científica pessoal é dada pela equação 1

$$Rep_i = \sum_{j=1}^{n} \frac{v_{ji} \times w_j}{\max(v_j)} \quad (1)$$

onde i representa o investigador cuja reputação é calculada, $Repi$ é a reputação do investigador i, j representa os critérios de reputação, vji representa o valor do critério j do investigador i, w representa o peso do critério j e max (vj) representa o valor máximo do critério j.

A reputação é ≥ 0. O valor dos critérios é extraído de uma base de dados e o seu valor normalizado varia entre [0, 1]. A soma do peso dos critérios varia, de igual forma, entre [0, 1].

3.3 Definição dos critérios de reputação

A tabela 2 apresenta os 20 critérios considerados para a métrica de reputação científica proposta, agrupados em quatro categorias: formação e percurso profissional, produtividade, projetos e reconhecimento. É complementada por uma descrição dos critérios e pelas fontes de onde os dados dos critérios devem ser extraídos.

Os critérios resultam do entendimento do fenómeno em estudo, da literatura consultada e, importante mencionar, dos dados disponíveis a aceder e extrair.

Os dados advêm do ORCID, uma organização cuja missão é fornecer um identificador único àqueles que estão envolvidos em atividades de investigação, académicas e de inovação.

A razão desta seleção prende-se com a diversidade de dados, com o caráter global do projeto e, sobretudo, do papel do ORCID na ciência. Atualmente, mais de 3 000 revistas científicas exigem os identificadores ORCID na submissão de ROs. A eLife, a Public Library of Science (PLOS) e The Royal Society comprometeram-se a solicitar ORCIDs nos fluxos de publicação (Harrison, 2017).

Os critérios da formação e percurso profissional respeitam a visão de Nicholas (2017), que argumenta que a jornada do investigador é fundamental de considerar. A educação recebida, a experiência profissional atual e passada, as qualificações profissionais adquiridas, as atividades à ordem de uma organização e a afiliação com associações científicas motivam os métodos de trabalhos de investigação, o pensamento e o espírito científico. Ademais, afeta a imagem que terceiros – quer pertençam à comunidade científica quer sejam externos a ela – formam do investigador.

Se a reputação é, em parte, o julgamento recebido pelos resultados de investigação publicados, então a produção científica é crucial para a reputação. Como o número total de publicações ignora a relevância do tipo de publicação (artigo em revista, conferência, livro etc.), a categoria produtividade encontra-se compartimentalizada. Em acrescento aos tipos de produção, a produtividade engloba igualmente as coautorias – visto afetarem a quantidade produzida e, possivelmente, a qualidade -, a propriedade intelectual – que resulta da investigação e pode ser indicador da qualidade desse output – e o v-índex.

Tabela 2 Critérios da reputação científica da métrica.

Categoria	Critério	Descrição do critério	Fonte
Formação e percurso profissional	Educação	Afiliação do investigador com universidades e respetivos graus académicos-	ORCID
	Emprego	Atividade profissional atual e passadas do investigador.	ORCID
	Qualificações	Prova da aquisição de competências como certificados e formação profissional.	ORCID
	Serviços	Atividades realizadas ao serviço de uma organização.	ORCID
	Associações membro	Afiliação do investigador com organizações ou sociedades enquanto membro.	ORCID
Produtividade	Artigos em revista	Publicações em revistas.	ORCID
	Artigos em conferência	Publicações de conferências.	ORCID
	Livros	Publicação de livros.	ORCID
	Capítulos de livros	Escrita de um ou mais capítulo de livros publicados em colaboração.	ORCID
	Livros editados	Publicações de livros editados parcial ou totalmente.	ORCID
	Manuais	Publicações de manuais escolares ou académicos.	ORCID
	Relatórios	Publicação de relatórios.	ORCID
	Peer reviews	Publicação de revisão de pares.	ORCID
	Coautorias	ROS feitos em colaboração.	ORCID
	V-índex	Valor do h-índex ajustado ao ano de publicação.	ORCID

O v-índex surge como uma expansão do h-índex para ultrapassar esse problema (Vaidya, 2005). O v-índex é o h-índex ajustado ao ano de publicação: $v = h - \frac{índex}{p(y_{this} - y_0)}$, onde y_{this} é o ano atual e y_0 é

o ano da primeira publicação. O v-índex integra o número de citações, motivo pelo qual o indicador não consta na lista de critérios.

Projetos compõe dois critérios: o financiamento e os recursos de investigação obtidos. No Reino Unido, o REF determina o financiamento futuro dos institutos. O financiamento é, de certa forma, uma medida de sucesso, como também afeta o processo da investigação. À semelhança, as bolsas de investigação e os recursos de investigação são um aspeto vital na ciência. Nessa medida, ambos os critérios respeitam a definição de reputação e por isso são contemplados.

Os prémios e distinções cômpares, e ofertas de trabalho procedem de avaliações positivas sobre os ROs por parte da comunidade científica e/ou sociedade em geral.

3.4 Definição dos pesos dos critérios de reputação

Os pesos dos critérios da fórmula de reputação $Repi$, são determinados pelos especialistas das áreas das Ciências da Computação, Informática, Sistemas de Informação e similares através de um estudo Delphi-FUCOM. É através deste estudo, em formato de questionário, que a métrica de reputação científica proposta é validada.

Durante o estudo, será pedido aos especialistas que façam uma comparação $n-1$ dos critérios de reputação por mim propostos, obtendo assim uma ordenação dos mesmos. Contudo, olhando para a equação 1 e levando em consideração o capítulo anterior, sabe-se que o valor dos critérios de cada investigador é obtido com o ORCID, mas o peso dos critérios e a reputação dos investigadores são 2 incógnitas.

Para calcular a reputação, é necessário conhecer o peso de cada critério, mas para fazer uma comparação $n-1$ dos critérios é necessário ter uma ordenação inicial dos critérios, que se obtém conhecendo a reputação do h-índex. Para resolver este problema, num primeiro momento, assume-se que a reputação do investigador i é igual ao seu h-índex. Esta decisão é motivada pelo facto de o h-índex, como visto na revisão de literatura à reputação científica, é a métrica mais aceite e empregue para determinar o impacto do trabalho de um investigador. Outra vantagem desta decisão é que, posteriormente, a solução final dos pesos dos critérios pode ser comparada com a solução obtida com o uso do h-índex. Será possível calcular a correlação entre os critérios e testar hipóteses relativas à fiabilidade do h-índice enquanto métrica da reputação científica.

3.4.1 O processo

Para obter o peso dos critérios utilizando o h-índex aplica-se o algoritmo Simplex LP. Com esse objetivo, construiu-se antecipadamente um modelo de programação linear *(linear programming, LP)*, composto por (Alvelos, 2009):

- Variáveis de decisão - traduzem matematicamente a decisão a tomar para o problema ser resolvido. Geralmente são valores incógnitos que têm de ser determinados.
- Restrições - limitam os valores que as variáveis podem tomar, por exemplo, a quantidade mínima ou máxima. Representam as alternativas de decisão possíveis.
- Função objetivo - indica o valor-objetivo otimizado. A função maximiza ou minimiza um valor numérico.

As variáveis de decisão correspondem, neste caso, ao peso de cada critério de reputação. A forma de resolver a equação é através da minimização da soma dos desvios absolutos. O método consiste em minimizar a soma incorrida dos desvios dos objetivos. Isto é, do desvio entre o valor estimado (da reputação, calculado com o h-índex) e o valor observado (computado pelo algoritmo).

Porque não se sabe se a solução (minimização da soma) irá sub ou sobre-satisfazer os objetivos, definem-se as variáveis auxiliares s_j e t_j . s_j representa a quantidade pela qual o objetivo é excedido numericamente. t_j representa a quantidade deficitária entre o objetivo e o valor numérico estimado. Note-se que ou apenas uma das variáveis auxiliares tem valor positivo ou ambas têm valor nulo (o valor estimado é igual ao valor observado).

Com o objetivo de minimizar a soma dos desvios absolutos, o modelo é:

$$Min\ Z_i = \sum_{j=1}^{n} (s_j + t_j) \tag{2}$$

Sujeito a:

$$Rep_i = \sum_{j=1}^{n} \frac{v_j \times w_j}{\max(v_j)} + (s_j + t_j)$$

$$s_j \geq 0, t_j \geq 0, w \geq 0$$

$$w\ livre \tag{3}$$

Para aplicar o modelo apresentado, testar e validar a métrica proposta, são necessários dados de de investigadores das áreas de Ciências da Computação, Sistemas de Informação e afins. O processo de identificação da população de investigadores é descrito no capítulo seguinte.

Em adição, a validação da métrica e obtenção dos pesos dos critérios é definido de acordo com o julgamento de especialistas das áreas científicas mencionadas através de uma abordagem híbrida do Método Delphi com o *Full Consistency Method* (FUCOM).

O processo proposto foi submetido junto da Comissão de Ética para a Investigação em Ciências Sociais e Humanas (CEICSH) da Universidade do Minho, e aprovado pelos seus membros (anexo I).

3.4.2 População de investigadores

Para identificar os investigadores a quem o modelo matemático do problema é aplicado, e como não existe uma diretoria ou base de dados completa da população universal de investigadores, foi utilizado um *web crawler* na página Investigadores do website Portal da Inovação[5], gerido e operado pela Agência da Nacional de Inovação[6]. Essa página web contém uma lista de investigadores alimentada pelo CiênciaVitae.

Um total de 386 investigadores com Ciências da Computação como domínios de atuação na plataforma CiênciaVitae e ORCID – para que os dados possam ser extraídos – foram identificados.

Anteriormente a esta população de investigadores foi identificada uma outra, que se revelou um percalço, motivos pelos quais se optou pela lista já descrita.

Numa 1ª versão, a população de investigadores correspondia aos investigadores de Ciências da Computação enumerados na lista *Highly Cited Researchers* 2020 da Clarivate[7].

A lista de 2020 contém cerca de 3 900 investigadores, distribuídos em 21 áreas das ciências e ciências sociais. Ela é composta por artigos altamente citados – artigos cuja contagem de citações situa um investigador no top 1% numa área de investigação durante o período de 2009-2019 – em revistas indexadas na Web of Science Core Collection™.

Da lista da Clarivate, foram selecionados todos os investigadores influentes na área da ciência da computação e associados a um id ORCID. De 125 investigadores de computação, 79 possuíam um

[5] https://www.portaldainovacao.pt/Portal_Inovacao/Researchers.aspx

[6] https://www.ani.pt/

[7] https://clarivate.com/webofsciencegroup/highly-cited-researchers-2020-executive-summary/

ORCID, sendo essa a população. No entanto, extraídos os dados, notou-se o problema de escassez de dados, o que torna esta população inadequada. Dos 20 critérios, 9 deles não tinham dados de nenhum dos investigadores, ou seja, eram dados vazios.

O obstáculo foi remediado pela procura e encontro da nova lista de investigadores do domínico científico pretendido.

3.4.3 Método Delphi

O Método Delphi é um processo estruturado e iterativo que solicita as opiniões de um grupo (painel) de especialistas (participantes) em relação a um problema, tópico ou tarefa. Geralmente, os especialistas são submetidos a uma série de questionários sequenciais (rondas), intercalados com informação condensada e feedback controlado (Ab Latif et al., 2016; Alarabiat & Ramos, 2019; Hasson et al., 2000; Rowe & Wright, 1999).

Devido à flexibilidade do Método (Hasson et al., 2000), os processos adotados para o implementar variam conforme os estudos. Todavia, pode-se sintetizar o Delphi em 5 caraterísticas determinantes:

1. Os participantes são especialistas na sua área (Ab Latif et al., 2016; Alarabiat & Ramos, 2019; Hasson et al., 2000).
2. Os participantes assumem uma identidade anónima (Ab Latif et al., 2016; Alarabiat & Ramos, 2019; Hasson et al., 2000; Rowe & Wright, 1999).
3. Os estudos de Delphi são executados em iterações (Ab Latif et al., 2016; Alarabiat & Ramos, 2019; Hasson et al., 2000; Rowe & Wright, 1999).
4. Os participantes recebem feedback ao longo do processo (Ab Latif et al., 2016; Alarabiat & Ramos, 2019; Hasson et al., 2000; Rowe & Wright, 1999).
5. As respostas dos participantes devem ser estatisticamente agrupadas (Hasson et al., 2000; Rowe & Wright, 1999).

Os participantes/especialistas são fulcrais ao Método Delphi. Tanto a qualidade das respostas como a qualidade dos resultados do estudo Delphi pende deles (Alarabiat & Ramos, 2019). Por esse motivo, selecionar os participantes/especialistas é uma etapa fundamental do processo.

Comumente, os especialistas são indivíduos informados com conhecimento e experiência pessoal sobre o problema, tópico ou tarefa a ser investigada (Hasson et al., 2000).

A anonimidade dos especialistas permite que, ao longo das rondas (iterações), estes expressem e alterem as suas opiniões sem pressão social ou medo de serem julgados (Ab Latif et al., 2016; Rowe & Wright, 1999).

O Método Delphi clássico é composto por quatro rondas, porém há evidência de que duas ou três é preferível. Não existe, no entanto, um entendimento incontestável. Há quem argumente que o número de rondas está relacionado com o alcance de um consenso aceitável entre os especialistas. Alguns sugerem um consenso de 51%, outros de 70% e ainda de 80%. A relevância do uso de uma percentagem também já foi questionada (Hasson et al., 2000). Assim, os estudos Delphi tendem a terminar quando se sente que se chegou a um consenso satisfatório ou quando a mudança das opiniões do especialista entre rondas sucessivas é mínima, mesmo que um consenso significativo não tenha sido atingido. Para além da estabilidade das respostas, o nível de concordância pode ser um indicador fiável (Alarabiat & Ramos, 2019; Hasson et al., 2000).

Entre cada ronda, os participantes/especialistas recebem feedback controlado sobre as respostas dos colegas na forma de um sumário estatístico de, por exemplo, tendências centrais como a média, mediana (Hasson et al., 2000; Rowe & Wright, 1999) e níveis de dispersão como o desvio padrão e o intervalo interquartil (Hasson et al., 2000). Supletivamente, informação como o raciocínio dos especialistas nas respostas pode ser partilhada.

Concluídas as rondas e reunidas os dados das respostas, as opiniões são agregadas e o julgamento conjunto equivale à média ou mediana estatística do grupo de especialistas.

3.4.4 Full Consistency Method

O *Full Consistency Method* é um método multicritério de resolução de problemas, a utilizar para determinar o peso dos critérios num problema multicritério de tomada de decisão (Pamučar et al., 2018). É uma abordagem subjetiva em que especialistas dão a sua opinião quanto à significância dos critérios, de acordo com as suas preferências.

O FUCOM baseia-se na comparação de critérios em pares, exceto apenas obriga à comparação $n - 1$. O modelo é validado através da otimização de uma função cujo objetivo é a minimização do desvio da consistência total da comparação. Uma outra vantagem é que facilita a consistência dos pesos dos critérios com a satisfação de condições de transitividade (restrições).

O procedimento para o FUCOM é o que se segue.

Passo 1. Os critérios $C = \{C_1, C_2, \dots, C_n\}$ são ordenados por ordem decrescente de importância (representado por k na fórmula 4), ou seja, do critério que se espera que tenha um coeficiente de maior peso maior para o critério que se espera que tenha um coeficiente de menor peso.

$$C_{j(1)} > C_{j(2)} > \cdots > C_{j(k)} \tag{4}$$

Passo 2. A prioridade comparativa dos critérios ordenados é determinada (equação 5). Representa ela a significância $\frac{\varphi k}{k+1}$ que o critério de ranking superior $C_{j(k)}$ tem sobre o critério de ranking inferior $C_{j(k+1)}$:

$$\Phi = \varphi\frac{1}{2}, \varphi\frac{2}{3}, \dots, \varphi\frac{k}{(k+1)} \tag{5}$$

A prioridade comparativa pode ser definida com base no julgamento dos participantes/especialistas, ou com base numa escala predefinida, os especialistas comparam os critérios e decidem a prioridade individual de cada.

Passo 3. Os valores finais dos coeficientes de peso $(w_1, w_2, \dots, w_n)^T$ são computados pelo seguinte modelo matemático:

$$\begin{gathered} \min \chi \\ s.t. \\ \left|\frac{w_{k(k)}}{w_{j(k+1)}} - \frac{\varphi k}{k+1}\right| \leq \chi, \forall j \\ \left|\frac{w_{k(k)}}{w_{k(k+2)}} - \frac{\varphi k}{k+1} \otimes \frac{\varphi(k+1)}{(k+2)}\right| \leq \chi, \forall j \\ \sum_{j=1}^{n} w_j = 1, \forall j \\ wj \geq 0, \forall j \end{gathered} \tag{6}$$

O modelo tem duas restrições, as quais devem ser respeitas: (1) o rácio dos coeficientes de peso é igual à prioridade comparativa dos pesos determinada no passo 2, (2) o valor dos coeficientes respeita transitividade. A consistência total é atingida quando a transitividade é completamente respeitada.

Ao resolver o modelo do problema, obtêm-se os valores dos pesos dos critérios.

O processo para a definição dos pesos dos critérios de reputação até agora apresentado reflete a decisão final de como proceder. Num momento anterior, uma outra abordagem foi ponderada: o *Fuzzy Delphi Method* (FDM). Nesta iteração, não era o peso dos critérios que seria definido pelos especialistas, mas antes a variável reputação. Quer isto dizer que, na equação 1, a incógnita a ser determina pelos especialistas não é o peso w_j e sim Rep_i . Segue-se a explicação da abordagem.

Embora amplamente aplicado (Chang et al., 2011), o Método Delphi tradicional é criticado pelo processo moroso, pela necessidade de inquéritos repetitivos, pela despesa, e pela diminuição da taxa de resposta à medida que os inquéritos se prolongam (Ab Latif et al., 2016; Ishikawa et al., 1993).

Ao fundir a teoria *Fuzzy* e o Método Delphi, consegue-se manter um resultado com mérito semelhante ao Método Delphi e, em simultâneo, reduzir o número de repetições do inquérito e os custos associados (Ishikawa et al., 1993). Também o FDM já foi adaptado para resolver as divergências de consenso, combinando-o com variáveis linguísticas (Chang et al., 2011; Mohamed Yusoff et al., 2021; Smarandache et al., 2020). Uma variável linguística contém valores que são palavras, e o significado dessas palavras são conjuntos *fuzzy* num determinado universo (Borovička, 2014).

Numa primeira etapa, seria pedido aos especialistas para determinarem a importância de cada critério e, de seguida, que respondessem a um conjunto de perguntas associadas e incidentes em investigadores anónimos – uma amostra da população de investigadores, retirada da população de investigadores do conjunto de dados. As respostas seriam restritas a uma escala Likert de 7 pontos, de *Extremamente sem importância* a *Extremamente importante*. Daí resultaria o valor de reputação, que ajudaria a estipular o peso dos critérios da métrica.

Originalmente, a reputação era também dada pela equação 1, mas o processo para a determinação dos pesos era diferente, como já mencionado em detalhe. Inicialmente, os especialistas determinariam diretamente o peso dos critérios, mas a ideia era selecionar um conjunto de investigadores (uma amostra), desse conjunto selecionar um certo número de ROs, e pedir aos especialistas para os avaliarem com a ajuda de um guia com critérios. Os critérios seriam qualitativos. Os valores de reputação calculados através deste modo seriam utilizados para calcular pesos por integração de subjetividade.

Numa segunda etapa, os especialistas responderiam ao inquérito expressando, linguisticamente, a relevância dos critérios, numa escala Likert de 7 pontos. A escala seria posteriormente traduzida matematicamente. Um dos requisitos do FDM é transformar as variações linguísticas, neste caso, a escala Likert, em números *fuzzy* triangulares (Chang et al., 2011; Yusoff et al., 2021). Um número triangular *fuzzy* pode ser formalmente escrito da seguinte forma: $\tilde{F} = (a, b, c)$ (Borovička, 2014).

Para melhor compreensão da transformação das respostas na escala Likert em números *fuzzy* triangulares, a tabela 3 mostra a correspondência entre a variação linguística e a escala *fuzzy*.

Tabela 3 Correspondência entre a variação linguística e a escala fuzzy*.*

Variável linguística	**Escala *fuzzy***
Concordo totalmente	(0.9, 1.0, 1.0)
Concordo	(0.7, 0.9, 1.0)
Concordo de alguma forma	(0.5, 0.7, 0.9
Não concord nem discord	(0.3, 0.5, 0.7)
Discordo de alguma forma	(0.1, 0.3, 0.5)
Discord	(0.0, 0.1, 0.3)
Discordo totalmente	(0.0, 0.0, 0.1)

Fonte: Yusoff et al. (2021).

Numa terceira fase, para cada especialista, computar-se-ia a distância entre a classificação média de r_{ij} (a classificação do investigador i no critério Cj) e X, e a distância entre a média de w_{ij} (o peso atribuído ao citério j pelo especialista i) e Y. A distância entre dois números *fuzzy* é dada por:

$$d(F_1, F2) = \sqrt{\frac{1}{K}\left[(m_1 - n_1)^2 + (m_2 - n_2)^2 + (m_3 - n_3)^2\right]} \qquad (7)$$

Se a distância d entre a média e a avaliação dos especialistas for ≤ 0,2, então considera-se que os especialistas chegaram a um consenso (Cheng & Lin, 2002). Caso tal se verifique, pode-se passar à etapa seguinte, caso tal não se verifique, deve-se informar cada um dos especialistas sobre os resultados e pedir uma explicação dos pesos atribuídos aos critérios. É necessário realizar mais uma ronda do questionário (Smarandache et al., 2020).

A quarta etapa consiste em agregar as avaliações *fuzzy*.

Na quinta etapa, para cada investigador avaliado pelos especialistas, a avaliação *fuzzy* é *defuzzified*. A segunda condição para o número *fuzzy* triangular é definir uma percentagem que, quando excedida, significa que o consenso entre o grupo de especialistas deve ser aceite. No Método Delphi tradicional, essa percentagem corresponde a 75%. No processo de *defuzzificação*, se o valor da pontuação *fuzzy* A for ≥ 0,5, então o valor médio é aceite. Se A for ≤ 0,5, então o valor médio é rejeitado.

No final, a opção de integrar o FDM e o método de cálculo de reputação subjetivo e qualitativo foi eliminada por esta abordagem restringir em demasia a definição de reputação científica pessoal. Ainda que inclua a subjetividade da reputação, dá uma imagem falsa da reputação do investigador.

Foi ainda ponderado optar pelo método *Analytic Hierarchy Process* (AHP), que incorpora a comparação de critérios pares. Por outras palavras, inclui a comparação de cada critério com outros

critérios/alternativas, em pares. Essa comparação é comumente feita numa matriz (tabela 4), que permite computar o peso dos critérios e o rácio da consistência.

Tabela 4 Exemplo da comparação a pares do método AHP.

	Critério 1	Critério 2	Critério 3	Critério 4
Critério 1				
Critério 2				
Critério 3				
Critério 4				

No entanto, a possibilidade foi colocada de parte considerando que, com 20 critérios, os especialistas terem de fazer um total de 190 comparações, o que poderia desincentiva-los e levá-los a abandonar a sua participação.

3.4.5 Recrutamento dos participantes do estudo Delphi-FUCOM

O processo de definição do peso dos critérios engloba um estudo Delphi, composto por um conjunto de questionários, sendo que o segundo questionário dependente dos resultados do primeiro e assim por diante.

Cimentado pela população de investigadores, o estudo destina-se a indivíduos que atuam nas áreas de conhecimento de Ciências da Computação, Informática, Sistemas de Informação e afins em centros de investigação sediados em Portugal.

Este estudo precisa de um mínimo de 5 participantes e não está restringido a um número máximo de participantes. Não existe consenso sobre o número de participantes para estudos Delphi, apenas recomendações e pareceres. Clayton (1997) refere que, para um grupo homogéneo de participantes, recomenda-se 15-30 participantes. Para um grupo heterogéneo de participantes, recomenda-se que haja 5-10 participantes (Delbecq et al., 1975). Ainda que alguns acreditem que para uma maior fiabilidade se exige um maior número de participantes, outros questionam esta posição. Para estes, não pode ser estabelecida uma relação causal entre a dimensão do painel e o rigor do consenso final. Significa isto que um grande painel de participantes não produz necessariamente melhores resultados. Akins et al. (2005) indicam que as características de resposta de um pequeno grupo de peritos numa área de conhecimento bem definida são estáveis à luz de uma amostragem aumentada. Grupos de participantes com compreensão geral numa área e/ou tópico são uma amostra fiável para desenvolver critérios fiáveis de apoio a tomadas de decisão.

Os participantes devem ser contactados via email, onde é esclarecido que não há obrigação em integrar o estudo. Em adição, para evitar influência ou coerção, o questionário do estudo deve explicitar que a participação é voluntária e pode terminar a qualquer momento, caso o participante assim o deseje.

Frequentemente, a seleção da amostra de participantes envolve técnicas de amostragem não probabilísticas. Os participantes não são selecionados aleatoriamente, pelo que a representatividade não é assegurada. Eles são selecionados com um objetivo: aplicar os seus conhecimentos de um determinado problema. Trata-se de uma amostragem propositada, que pressupõe que os conhecimentos de um investigador sobre a população podem ser utilizados para escolher os casos a incluir na amostra (Hasson et al., 2000).

No caso do projeto, os participantes são identificados/selecionados com base na amostragem propositada. Como critérios de inclusão, os participantes devem (i) ter, pelo menos, 5 anos de experiência profissional no domínio das Ciências da Computação, Informática, Sistemas de Informação e afins; (ii) exercer a sua atividade profissional em Portugal; e (iii) estar (ou terem estado) associados à academia (investigadores). Aqueles que não satisfazerem estes critérios não são contactados

3.4.6 Procedimento do estudo Delphi–FUCOM

Num primeiro momento, são contactados potenciais participantes a pedir a sua colaboração no estudo. Àqueles que aderirem é enviado um primeiro inquérito onde se explica o objetivo e contexto do projeto.

Através do inquérito recolhe-se informação pessoal (género, idade) e informação sobre a experiência profissional (anos de experiência profissional na área das Ciências da Computação, Informática, Sistemas de Informação e afins; ocupação profissional atual; experiência direta ou indireta em avaliação de académicos; investigadores ou pares).

Também através desse primeiro questionário, é recolhido o parecer dos participantes em relação a uma lista de critérios (n) a utilizar para o cálculo da reputação científica pessoal. Esse parecer resulta de uma comparação $n - 1$, ou seja, comparação em pares, dos critérios.

Do questionário resulta uma ordenação dos critérios da reputação científica, que será apresentada aos participantes entre o primeiro e o segundo questionário do estudo.

No segundo questionário, os participantes fazem novamente uma comparação em pares, mas os critérios que são comparados dependem da ordem obtida no primeiro questionário.

Caso os participantes assim o indiquem, serão contactados no futuro com os resultados da investigação após a sua publicação.

O estudo Delphi é composto por rondas e termina quando se sentir que se chegou a um consenso satisfatório (Alarabiat & Ramos, 2019). As rondas são compostas por questionários online, realizados em dias diferentes e submetidos online. Para cada questionário, os participantes devem submeter as suas respostas num prazo de uma semana. Entre o envio do primeiro questionário para preenchimento e a submissão do segundo questionário, o estudo deverá ter uma duração total de, aproximadamente, 3 semanas.

Não ocorrerá qualquer gravação vídeo ou áudio. No entanto, vão ser coletados dados e informação pessoal. A informação agregada será utilizada apenas para fins do projeto, nomeadamente, publicações relacionadas com a investigação.

3.4.7 Confidencialidade do estudo Delphi–FUCOM

O estudo é totalmente confidencial. Um participante não será capaz de saber quem são os restantes participantes pois eles são abordados individualmente e esse tipo de informação não será fornecida.

Entre o primeiro e segundo questionário, ao indicar o resultado das respostas do primeiro questionário, a informação será agregada, pelo que não será possível discernir identidades.

O questionário é feito com recurso à ferramenta Tally[8], pelo que as respostas são recolhidas através desse meio.

Os dados serão anonimizados e disponibilizados de forma aberta no repositório de dados da Universidade do Minho, de modo a que possam ser escrutinados, usados e reusados por outros investigadores.

3.4.8 Benefícios e riscos do estudo Delphi–FUCOM

Com o estudo pretende-se construir e validar uma métrica de cálculo da reputação científica. Visto a métrica dizer respeito aos participantes – que devem, ao longo da sua carreira, ter-se deparado diversas vezes com o conceito de reputação científica e talvez até já tenham sido sujeitos a avaliações da sua reputação enquanto profissionais –, estes podem ter interesse pessoal em integrar o estudo. Contudo, não existe um benefício direto e/ou imediato.

Um potencial entrave à participação é a disponibilidade de tempo necessário para completar o questionário. No entanto, visto o questionário ser online, o participante pode escolher o momento mais

[8] https://tally.so/

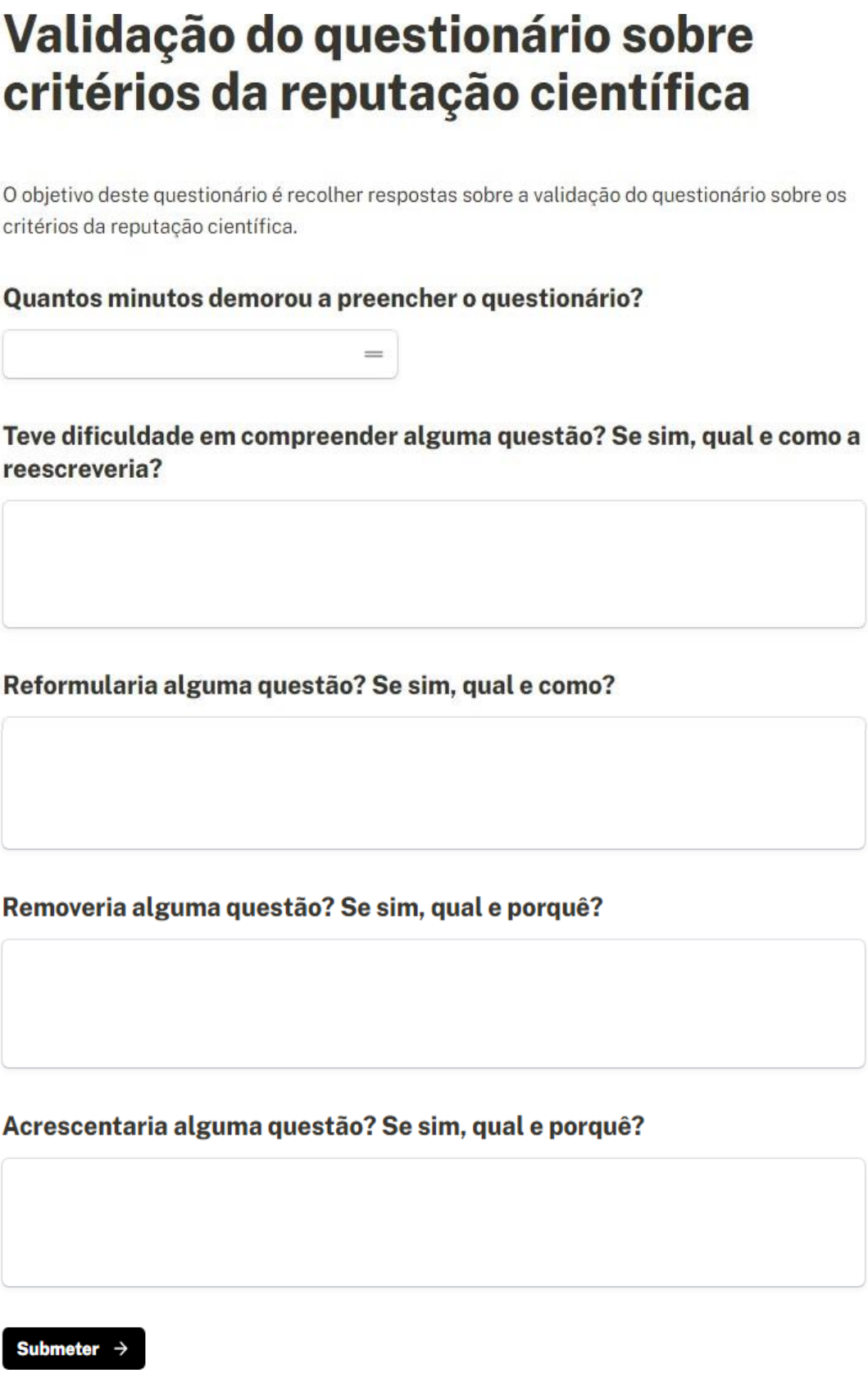
oportuno para responder. Esta é uma solução que pode diminuir a hesitação em disponibilizar o tempo necessário.

3.4.9 Validação do questionário do estudo através de um teste-piloto

O questionário que acompanha o estudo, ao qual se nomeou Critérios para cálculo da reputação científica (apêndice I), foi também aprovado pela CEICSH. Para provar a validade do questionário a enviar aos participantes, deve-se elaborar um teste-piloto para obter a aprovação do mesmo e determinar a necessidade de o redesenhar parcial ou totalmente.

Para esse efeito, o questionário deve ser enviado para um número reduzido de pessoas, juntamente com um segundo questionário, *Validação do questionário sobre critérios da reputação científica* (figura 2). O questionário de validação tem o propósito de recolher feedback quanto ao primeiro questionário, nomeadamente sobre o tempo de preenchimento, compreensão do enunciado e instruções, e adequação das questões colocadas.

Figura 2 Questionário Validação do questionário sobre critérios da reputação científica.



Validação do questionário sobre critérios da reputação científica

O objetivo deste questionário é recolher respostas sobre a validação do questionário sobre os critérios da reputação científica.

Quantos minutos demorou a preencher o questionário?

Teve dificuldade em compreender alguma questão? Se sim, qual e como a reescreveria?

Reformularia alguma questão? Se sim, qual e como?

Removeria alguma questão? Se sim, qual e porquê?

Acrescentaria alguma questão? Se sim, qual e porquê?

Submeter →

3.4.10 Obtenção dos pesos dos critérios e plano de validação da métrica

Para obter os pesos dos critérios de reputação e a validação da métrica, precisa-se do Excel (o semelhante), da ferramenta Solver do Excel e Tally, onde os questionários são estruturados, e da aplicação das equações do método FUCOM descritas em 3.4.4.

No Excel, depois de extraídos os dados sobre a população de investigadores para cada um dos 20 critérios, é aplicado o modelo de programação linear apresentado em Processo para a definição dos pesos dos critérios de reputação no Solver. A tabela 5 exibe uma amostra do resultado desse processo.

Tabela 5 Amostra dos dados recolhidos dos critérios de reputação.

orcid	journal articles	conference papers	books	book chapters	edited book	report
0000-0001-8144-4583	0,04	0,08	0,06	0,32	0,00	0,00
0000-0003-3685-0659	0,01	0,01	0,00	0,08	0,00	0,00
0000-0002-4163-905X	0,00	0,00	0,00	0,01	0,00	0,00
0000-0003-0682-2542	0,03	0,04	0,05	0,08	0,00	0,00
0000-0002-2308-6623	0,09	0,01	0,00	0,02	0,00	0,00
0000-0001-6018-7346	0,02	0,01	0,03	0,03	0,00	0,00
0000-0002-7900-9846	0,04	0,09	0,11	0,00	0,00	0,00
0000-0001-8150-9666	0,00	0,00	0,00	0,00	0,00	0,00
0000-0003-2858-1997	0,00	0,00	0,00	0,00	0,00	0,00
0000-0002-4144-8499	0,01	0,01	0,00	0,00	0,00	0,00
0000-0001-8060-5920	0,00	0,01	0,02	0,00	0,00	0,00
0000-0001-8803-0893	0,02	0,00	0,00	0,00	0,00	0,00
0000-0002-2255-8983	0,00	0,00	0,00	0,00	0,00	0,00
0000-0002-4721-6082	0,00	0,03	0,01	0,06	0,00	0,00
0000-0002-7085-1260	0,00	0,00	0,00	0,00	0,00	0,00

É de lembrar que, nesta fase, a reputação é calculada como se este valor fosse igual ao h-índex. O objetivo destes passos é obter uma ordenação dos critérios que sirva de base aos especialistas que preencherão o questionário do estudo Delphi-FUCOM.

A figura 3 demonstra como o modelo de programação da métrica foi construído no Excel. Cada linha corresponde a um investigador, sendo que a primeira de todas legenda as colunas. As linhas em branco correspondem ao valor normalizado dos critérios de cada investigador. As células de cor salmão na segunda linha representam os pesos dos critérios a serem calculados pelo Solver (a figura mostra o problema resolvido). Em conjunto com as duas colunas das variáveis auxiliares s_j e t_j , elas representam as variáveis de decisão. A coluna a verde possui a fórmula 19, e a célula azul possui a fórmula 18, ou seja, o objetivo de minimizar a soma dos desvios absolutos.

Figura 3 Modelo da métrica no Solver.

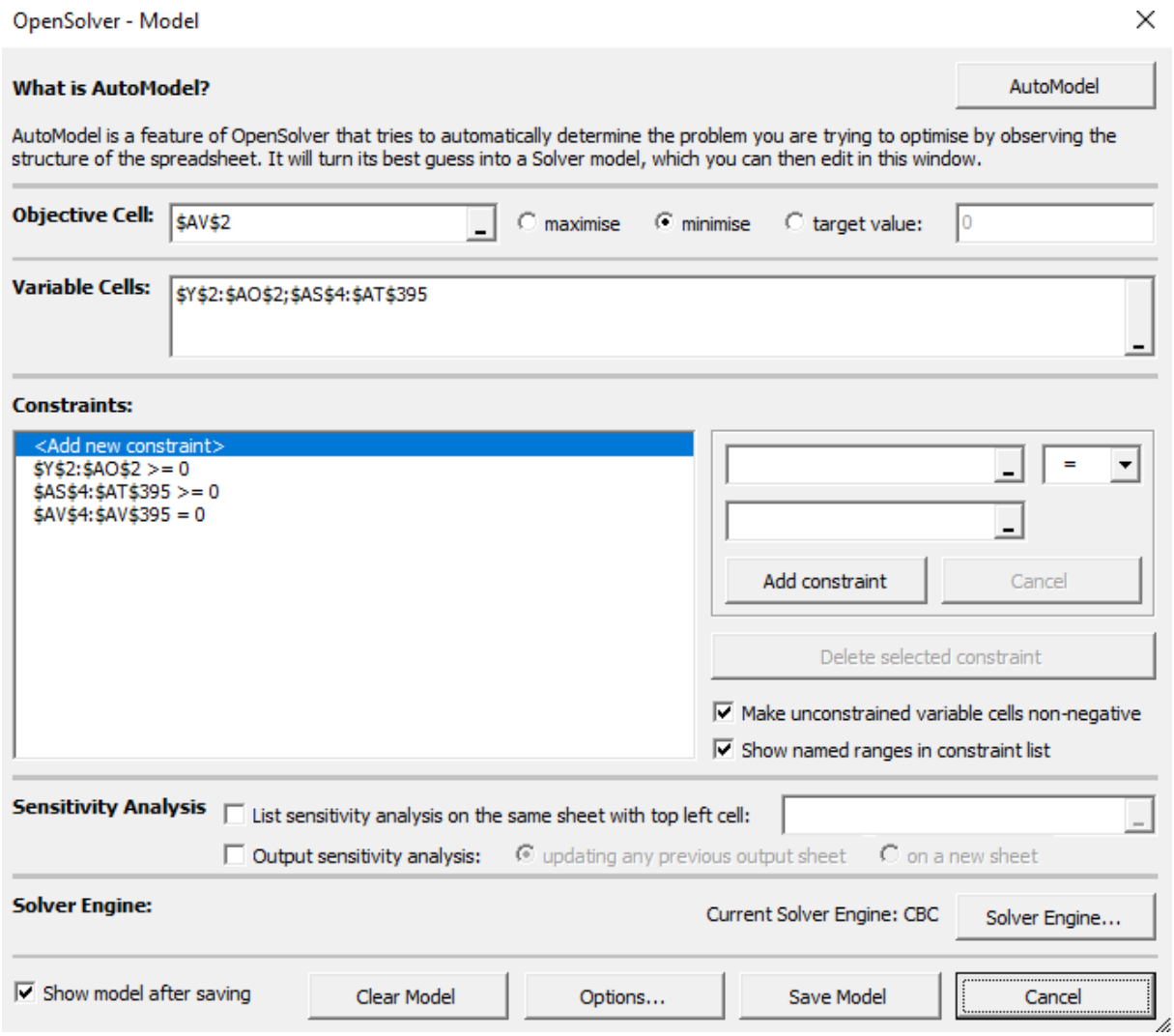
edited boc	report	manual	patent	distinctior	education	employme	qualificat	service	funding	peer revie	coauthor	v-index	invited positions	membership	research resource			
0	0	0	0,01324	0	0,00645	0,00431	0	0	0,04523	0,06673	0,36669	0,09451541	0%	0%	0%	0%		10,0305
																s	t	
0	0	0	0	0	0	0,00018	0	0	0	0	0,00684	2,9082E-05	0	0	0	0,03548	0	0%
0	0	0	0	0	0,00193	0,00108	0	0	0	0	0	0	0	0	0	0	0,00836	0%
0	0	0	0	0	0	0	0	0	0,00137	0	0	0	0	0	0	0	0,00221	0%
0	0	0	0	0	0,00129	0	0	0	0	0	0,00101	1,4541E-05	0	0	0	0,01392	0	0%
0	0	0	0	0	0	0	0	0	0,01508	0	0,24969	0,00018903	0	0	0	0	0,02765	0%
0	0	0	0	0	0,00258	0,00108	0	0	0,00685	0	0,0076	0,00011633	0	0	0	0,03641	0	0%
0	0	0	0	0	0	0	0	0	0	0	0,02254	7,2704E-05	0	0	0	0,04245	0	0%
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00171	0%
0	0	0	0,0012	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0012	0%
0	0	0	0	0	0,00193	0,00018	0	0	0	0	0	0	0	0	0	0	0,0074	0%
0	0	0	0	0	0,00193	0,00054	0	0	0	0	0,00431	0,00040714	0	0	0	0,02643	0	0%
0	0	0	0	0	0,00258	0,00036	0	0	0	0,00092	0	0	0	0	0	0	0,00427	0%
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2,1E-05	0%
0	0	0	0	0	0,00064	0,00018	0	0	0	0	0,00051	7,2704E-05	0	0	0	0,00339	0	0%
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%
0	0	0	0,0012	0	0,00258	0,00036	0	0	0	0	0,00456	4,3622E-05	0	0	0	0,00381	0	0%
0	0	0	0	0	0,00258	0,00054	0	0	0,01234	0,00015	0,00177	2,9082E-05	0	0	0	0	0,00746	0%
0	0	0	0	0	0,00129	0,00018	0	0	0,00822	0	0,00304	0,0003926	0	0	0	0,02611	0	0%
0	0	0	0	0	0,00064	0,00018	0	0	0	0	0	0	0	0	0	0	0,00167	0%
0	0	0	0	0	0,00516	0,00036	0	0	0,00137	0	0	0	0	0	0	0	0,00689	0%
0	0	0	0	0	0,00129	0,00018	0	0	0	0	0	0	0	0	0	0	0,00235	0%

Sabendo qual é a célula objetivo (célula azul), as variáveis de decisão (células de cor salmão), e as restrições a que o modelo tem de obedecer (secção 3.4.1), basta dar essas indicações no Solver, como demonstrado na figura 4.

Com a aplicação do algoritmo Simplex, obteve-se a seguinte ordenação dos critérios, por ordem decrescente, do mais importante para o menos importante para a reputação, segundo o h-índex: Artigos de conferência > co-autorias > v-índex > *peer reviews* > financiamento > livros > artigos de revista > propriedade intelectual > educação > emprego > capítulos de livros > livros editados > manuais > relatórios > prémios e outras distinções > qualificações > serviços > recursos de investigação = convites para cargos de trabalho = associações membro.

Obtida essa ordenação, foi redigido o primeiro inquérito do estudo Delphi-FUCOM, *Critérios para cálculo da reputação científic*a, a enviar aos participantes, após a validação do mesmo.

Figura 4 Preenchimento do Solver.



Obtida essa ordenação, foi redigido o primeiro inquérito do estudo Delphi-FUCOM, *Critérios para cálculo da reputação científica,* a enviar aos participantes, após a validação do mesmo.

Uma vez obtidas as respostas, deve-se aplicar as fórmulas presentes em *Full Consistency Method* recorrendo ao Excel, o que irá resultar numa nova ordenação dos critérios. De seguida, pegando no modelo do inquérito Critérios para cálculo da reputação científica, redige-se um outro, em que a única diferença é a apresentação dos critérios para os participantes efetuarem a comparação $n - 1$ destes. Posteriormente, às respostas do segundo questionário, devem ser novamente aplicadas as instruções e fórmulas da secção *Full Consistency Method*. Desta etapa deve resultar os pesos finais dos critérios de reputação, validados pela comunidade científica.

4. REVISÃO DE LITERATURA AOS SISTEMAS DE RECOMENDAÇÃO CIENTÍFICA

Nos capítulos que se seguem, *Metodologia* informa da metodologia na qual esta revisão literária se baseia; *Técnicas de recomendação* introduz e explica em pormenor as técnicas das quais os algoritmos de recomendação se servem para gerar recomendações, incluindo as principais limitações das técnicas e soluções propostas por investigadores para as ultrapassar; e *Avaliação dos sistemas de recomendação* apresenta os três tipos diferentes de avaliação que se podem aplicar aos sistemas de recomendação, nomeadamente avaliação offline, estudos de utilizador e avaliação online, e o procedimento para os realizar.

4.1 Metodologia

A elaboração da presente revisão de literatura é orientada pela declaração *Preferred Reporting Items for Systematic reviews and Meta-Analyses* (PRISMA) de Moher et al. (2009). Como o nome sugere, ela foi desenvolvida com o objetivo de auxiliar o processo de reportar revisões sistemáticas e meta análises.
Ela é composta por um fluxograma (figura 5), que representa visualmente o fluxo de informação ao longo das 4 fases do processo de revisão: 1) identificação, 2) triagem, 3) elegibilidade e 4) inclusão.
Na primeira fase, foram identificadas possíveis referências bibliográficas através do uso da base de dados Scopus[9]. A pesquisa efetuada pode ser replicada utilizando:

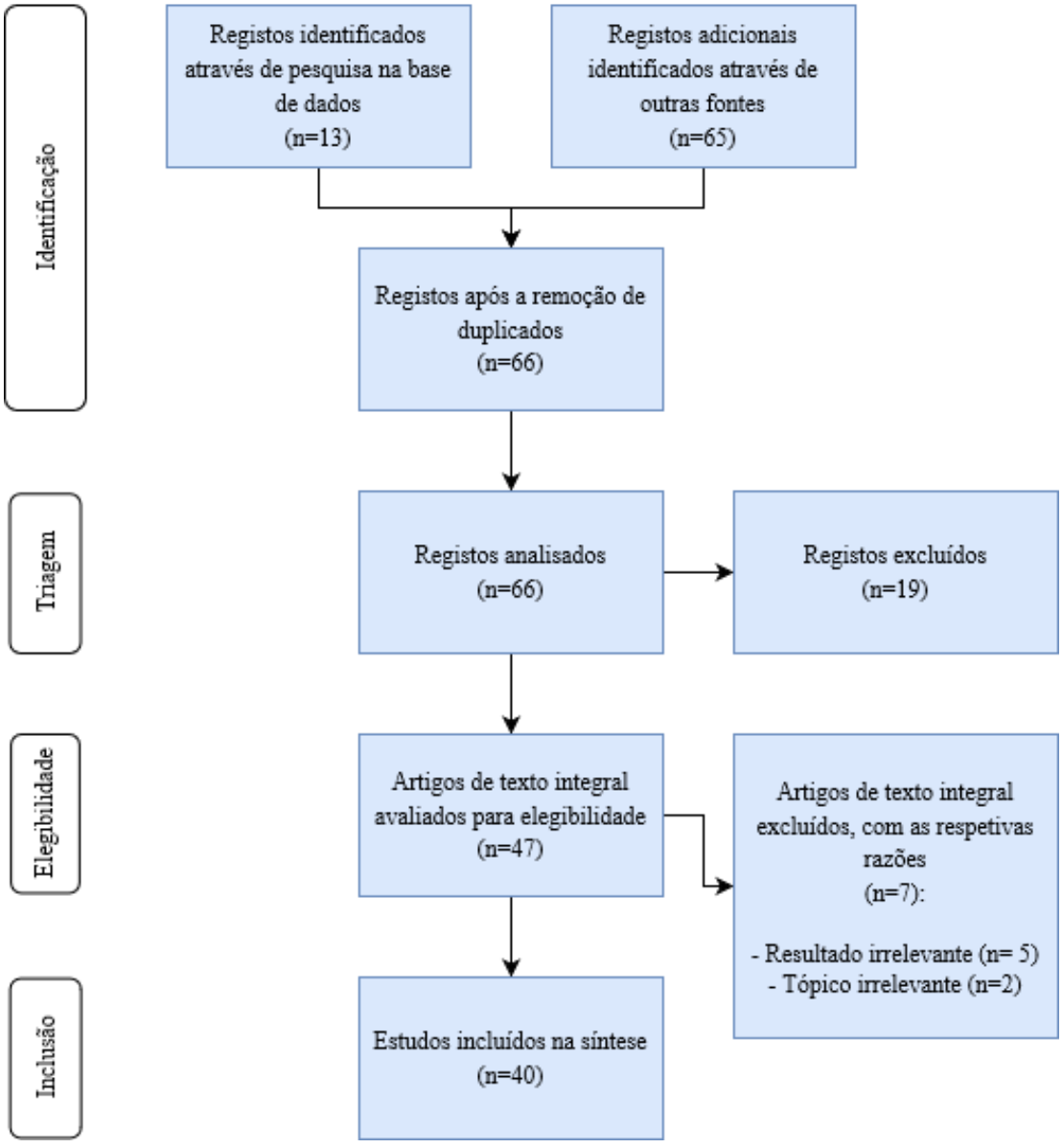
> [1] TITLE-ABS-KEY (''recommender system*'' OR ''recommendation system*'' AND (scholarly OR ''research-paper'' OR academic)) AND PUBYEAR > 2016 AND (LIMIT-TO (SUBJAREA, ''COMP'')) AND (LIMIT-TO (DOCTYPE, ''re'')) AND (LIMIT-TO (LANGUAGE, ''English''))

Como se pode observar, aos resultados de [1] foram impostas algumas condições.
Com o propósito de obter trabalhos de investigação recentes, impôs-se o critério de ano de publicação superior a 2016. Uma vez que o tema da revisão recai sobre a área de Ciências da Computação também os resultados foram filtrados para se produzir uma lista de registos mais relevante. Em acrescento, como

[9] https://scopus.com

ponto de partida, restringiu-se a pesquisa a revisões de literatura, com o intuito de consultar as suas referências e citações. Por fim, as publicações tinham de estar em língua inglesa.

Figura 5 Fluxograma da revisão de literatura aos sistemas de recomendação científica.



Fonte: Aaptado de Moher et al. (2009).

[1] gerou um total de 13 resultados, aos quais apenas se obteve acesso a 9 deles. Os restantes dependiam de uma subscrição paga. Houve tentativa de contacto com os respetivos autores, porém não se obteve resposta.

O número de registos identificados através do Scopus foi expandido com os processos de *backward* e *foward reference searching*. O primeiro envolve a identificação e análise das referências bibliográficas de um trabalho científico, A, a análise das referências identificadas em A, e por aí adiante, em cadeia. O

segundo envolve a identificação de trabalhos que citam os registos identificados, as referências bibliográficas desses registos e outros.

Durante a triagem, foi feita uma seleção preliminar de registos, com a adequação do título e resumo enquanto critérios. Nesta fase, 19 de 66 registos foram eliminados.

Os registos que passaram à fase seguinte, elegibilidade, foram lidos na íntegra, o que determinou a inclusão ou exclusão destes na revisão de literatura. Por motivos de irrelevância dos resultados dos registos ou dos tópicos abordados, 7 foram rejeitados.

No total, esta revisão de literatura contém 40 estudos.

4.2 Técnicas de recomendação

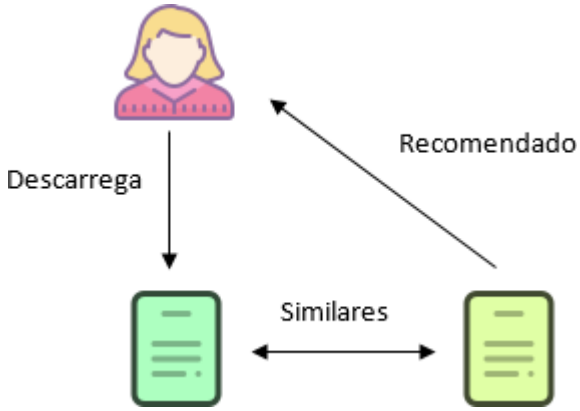
Este capítulo apresentas as 4 técnicas de recomendação que um SR pode adotar:

- *Content-Based* (CB) quando o algoritmo de recomendação do sistema utiliza a semelhança entre as características dos itens para gerar recomendações,
- *Collaborative-Filtering* (CF) quando o algoritmo de recomendação utiliza a semelhança entre as classificações que o utilizador dá aos itens para gerar recomendações,
- *Graph Based (GB)* quando o algoritmo de recomendação utiliza as relações), modeladas por gráficos, entre objetos (como utilizadores, itens e atributos),
- *Hybrid*, quando o algoritmo de recomendação utiliza uma combinação de diferentes técnicas de recomendação (por exemplo, *content-based* com *collaborative filtering*) para gerar recomendações.

4.2.1 Content-based

Um SR com filtragem baseada no conteúdo recomenda itens ao utilizador com base na semelhança entre os atributos dos itens por ele apreciados positivamente (figura 6) (Hanyurwimfura et al. 2015).

Figura 6 Ilustração do conceito de um SR content-based.



Um item é representado por um modelo de conteúdo que contém as suas caraterísticas, por exemplo, a área científica, título, resumo, palavras-chave, corpo de texto e autor de um artigo científico.

Os interesses, preferências e gostos do utilizador são extraídos a partir das caraterísticas dos itens com que ele interage, sob a forma de descarregamento, compra, citação ou outros.

Os itens com que o utilizador interage, denominados itens-alvo, são resumidos num perfil de item, que é depois comparado com os itens passíveis de serem recomendados, os chamados itens-candidatos.

Quanto maior o nível de similaridade entre o item-alvo e o item-candidato, maior a probabilidade de o item-candidato ser recomendado. Dito de outra forma, a técnica *content-based filtering* (CBF) recomenda itens similares aos que o utilizador apreciou no passado.

As caraterísticas dos itens podem ser extraídas em formato de palavras-chave ou de frases-chave (Hanyurwimfura et al. 2015; Ferrara F., Pudota N., Tasso C. 2011; Belém et al. 2014). Três conceitos estão subjacentes:

- *Vector-Space Model* (VSM): modelo que representa informação textual sob a forma de vetores. As componentes de cada vetor representam ou a importância dos termos ou a ausência/presença dos termos num documento/corpus. A métrica *Term Frequency and Inverse Document Frequency* é uma das mais utilizadas para determinar a importância dos termos.
- *Term Frequency and Inverse Document Frequency* (TF–IDF): TF (equação 8) indica a frequência do termo i no documento j, ou seja, o número de vezes que uma palavra aparece num dado documento.

$$TF(i,j) = \frac{Número\ total\ de\ vezes\ que\ o\ termo\ i\ aparece\ no\ documento\ j}{Número\ total\ de\ termos\ no\ documento\ j} \qquad (8)$$

Todavia, o TF ignora a importância das palavras. Por exemplo, os artigos "o(s)" e "uma(s)", e os conetores "mas" e "também" aparecem com regularidade na escrita, contudo estão

desprovidos de real valor. É o IDF (equação 9) que normaliza o peso dos termos de elevada frequência.

$$IDF(i) = \log_{10} \frac{Número\ total\ de\ documentos}{Número\ total\ de\ documentos\ com\ o\ termo\ i} \quad (8)$$

A multiplicação do TF pelo IDF (equação 10) determina a importância do termo i no documento j (Jiang et al. 2012).

$$w(i,j) = TF(i,j) \times IDF(i) \quad (9)$$

- *Cosine* (cos) *similarity* (sim): mede a semelhança entre dois vetores calculando o cosseno do ângulo entre ambos (Philip et al. 2014). Imagine-se que se quer determinar o nível de similaridade entre o artigo de investigação *u* e o artigo de investigação *v*. A sua computação é dada pela equação 11

$$cos = sim(\vec{u}, \vec{v}) = \frac{\vec{U} \times \vec{V}}{\|\vec{U}\| \times \|\vec{V}\|} = \frac{\sum_{i=1}^{n} U_i \times V_i}{\sqrt{\sum_{i=1}^{n} U_i^2} \times \sqrt{\sum_{i=1}^{n} V_i^2}} \quad (10)$$

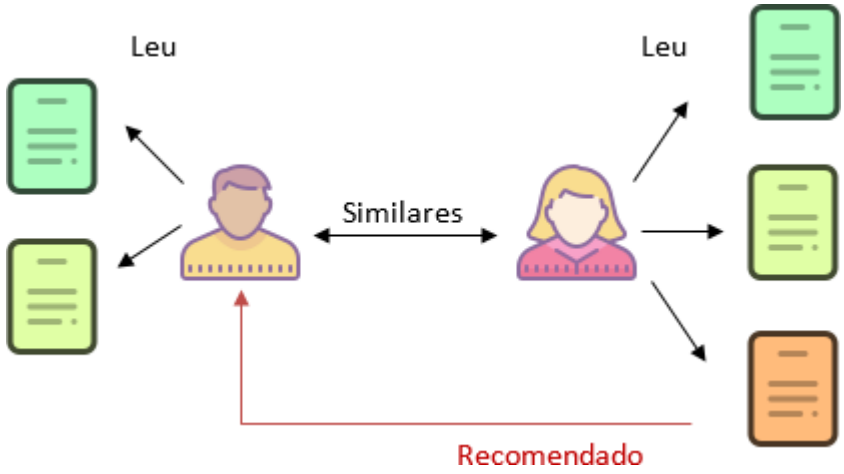
O *cos* calcula a similaridade entre os vetores dos itens e os vetores dos perfis de utilizador, ou entre dois vetores de itens. Philip et al. (2014) utilizam o *cos* para determinar as similaridades entre o corpus de documentos e a pesquisa do utilizador ativo. Sugiyama e Kan (2010) utilizam-no para estimar os itens mais semelhantes ao item-alvo, valendo-se das citações e referências.

O valor do *cos* varia entre -1 e 1. Quanto mais próximo de 1, maior a similaridade entre os vetores, e menor o ângulo entre eles. Se dois vetores fazem um ângulo de 0°, o cosseno é de1 e a semelhança de 100%. O inverso aplica-se quanto mais próximo o cosseno estiver de -1.

4.2.2 Collaborative filtering

Os sistemas de recomendação de *collaborative filtering* (filtragem colaborativa) baseiam-se no pressuposto de que utilizadores com gostos semelhantes no passado tendem a ter gostos semelhantes no futuro (Taneja e Arora 2018; Tang et al., 2021; Alfarhood e Cheng 2019 - 2019). Por exemplo, se investigador A e investigador B classificaram os artigos X e Y de forma semelhante no passado, então, A e B irão apreciar o artigo U de modo similar no futuro (figura 7).

Figura 7 Ilustração do conceito de collaborative filtering.



A relação entre os utilizadores é estabelecida pela semelhança entre as classificações que cada um deu aos mesmos itens. Os algoritmos recomendam, portanto, itens ao utilizador ativo que utilizadores semelhantes tenham gostado (Son e Kim 2018).

As classificações explícitas e as classificações implícitas são os dois principais processos para construir o perfil de utilizador (Wang et al. 2006).

Nas classificações explicitas, o utilizador fornece input ao SR expressamente. Elas assumem forma numérica (exemplo: de 1 a 5 estrelas), binária (exemplo: gosto / não gosto), ou unária (exemplo: botão "gosto", sem opção de dar "não gosto") (Jugovac e Jannach 2017). As preferências do utilizador podem ainda ser obtidas através de formulários, onde os próprios utilizadores as indicam, ou através de sistemas baseados em diálogo, onde os itens são apresentados ao utilizador de forma conversacional (Jugovac e Jannach 2017).

Nas classificações implícitas, os gostos, preferências e interesses do utilizador são inferidos a partir de ações, tais como o download, cliques em botões e links, tempo de leitura de uma página web, citação de artigos científicos e *social bookmarking*,(Amami et al., 2017; Shahid et al. 2020). Neste caso, são a monitorização e os registos de atividade que enriquecem o perfil do utilizador (Shahid et al. 2020).

Geralmente, as classificações dadas pelos utilizadores aos itens são representadas numa matriz de pontuação, onde cada linha corresponde a um histórico de feedback de um utilizador específico (Son e Kim 2018). A tabela 6 exemplifica uma matriz de classificação explícita e a tabela 7 exemplifica uma matriz de classificação implícita.

Tabela 6 Matriz de classificação explícita numa escala de 1 a 5.

	Item1	I2	I3	I4
Utilizador1		4	2	
U2	1		3	
U3	5	4	2	
U4	3		3	5

Tabela 7 Matriz de classificação implícita.

	I1	I2	I3	I4
U1		1		1
U2	1	1		
U3				1
U4	1		1	

De acordo com Amami et al. (2017) e Taneja e Arora (2018), a *collaborative filtering* (CF) divide-se em duas classes:

1. baseada em memória (ou vizinhança),
2. baseada em modelos.

Na **CF baseada em memória**, o algoritmo de recomendação encontra a relação entre utilizadores, ou itens similares, através de uma matriz de classificações (Sakib et al. 2020), e recomenda ao utilizador ativo os itens que os seus utilizadores vizinhos, ou utilizadores similares, classificaram de forma elevada (Shahid et al. 2020; Bai et al. 2019). A abordagem baseada em memória desdobra-se também em duas categorias (Boussaadi et al. 2020 - 2020):

- *Baseada no utilizador*: primeiro são determinados os utilizadores semelhantes ao utilizador ativo. Corresponde isto a estimar as semelhanças entre a linha na matriz de classificação do utilizador ativo com as linhas dos outros utilizadores (Resnick et al. 1994). De seguida, os vizinhos do utilizador ativo são selecionados com base nas similaridades entre eles (tabela 8). As classificações do utilizador ativo são previstas de acordo com a média das classificações ponderadas conhecidas do item-candidato pelos utilizadores semelhantes (Wang et al. 2008). São assim preenchidas as células vazias da matriz com uma pontuação de previsão. Por último, os itens são apresentados ao utilizador numa lista top-N, onde N denota um valor numérico.

Tabela 8 Ilustração do conceito de collaborative filtering *baseada no utilizador.*

	Item8	Similaridade	Média de classificações
João	?	1	4
Utilizador1	3	0,91	2,4
Utilizador 2	5	0,83	3,8

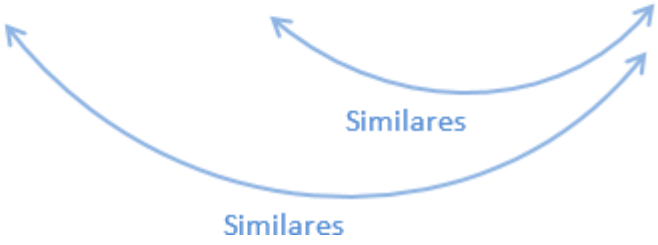
Similares

Destaca-se duas formas de calcular os vizinhos mais próximos (*nearest neighbors*) do utilizador ativo: *k-nearest neighbors* e *setting threshold*. Com o método *k-nearest neighbors*, os primeiros k utilizadores mais semelhantes ao utilizador ativo são selecionados. No método *setting threshold*, é definido um limite numérico. Quando a semelhança entre o utilizador *v* e o utilizador ativo *u* é maior que o limiar, o utilizador *v* é selecionado como sendo um dos vizinhos mais próximos.

- *Baseada no item*: primeiro são determinados os itens semelhantes ao item-alvo (tabela 9). Tal como na CF baseada no utilizador, os itens-alvo são normalmente calculados com os métodos *k-nearest neighbors* e *setting threshold*. Consiste em prever a classificação do utilizador ativo para um item-candidato, com base nas suas avaliações a itens semelhantes ao item- candidato (Adomavicius e Tuzhilin 2005). A classificação desconhecida é prevista através da média das classificações ponderadas conhecidas dos itens semelhantes do utilizador ativo (Wang et al. 2008).

Tabela 9 Ilustração do conceito de collaborative filtering *baseada em itens.*

	Item1	Item2	Item3	Item4	Item5	Item6	Item7
Maria	2	1	4	5	4	3	?
Utilizador1	3	1	4	2	3	4	3
Utilizador2	5	3	3	1	4	5	5
Utilizador3	1	5	1	3	2	4	2
Utilizador4	2	4	2	4	3	5	1



Quer para a CF baseada no utilizador quer para a CF baseada no item, a semelhança entre utilizadores ou itens é estimada através do *cos* e ou do coeficiente de correlação de Pearson (Taneja e Arora 2018; Boussaadi et al.).

Aplica-se o coeficiente de correlação Pearson (Resnick et al. 1994) quando se quer encontrar o grau de similaridade entre um par de itens ou um par de utilizadores. A equação 12 exemplifica o cálculo da similaridade ente utilizadores, onde $\overline{r_u}$ e $\overline{r_v}$ representam a média das classificações dos utilizadores u e v, respetivamente, com este coeficiente.

$$\mathrm{sim(u,v)} = \frac{\sum_{i \in (u.v)} \left(r_{u,i} - \overline{r_u}\right) \times \left(r_{v,i} - \overline{r_v}\right.}{\sqrt{\left(r_{u,i} - \overline{r_u}\right)^2} \times \sqrt{\left(r_{v,i} - \overline{r_v}\right)^2}} \qquad (12)$$

O valor do coeficiente de correlação Pearson varia entre -1 e +1, sendo que quanto mais próximo o valor estiver de +1, maior a correlação entre utilizadores e, por isso, maior a semelhança dos gostos. Por outro lado, quanto mais próximo o valor estiver de -1, menor a correlação entre utilizadores e maior a disparidade de gostos (Taneja e Arora 2018).

A principal desvantagem de um SR baseado em memória é a sua inadequação para lidar com um número elevado de utilizadores e itens

Na CF **baseada em modelos**, o sistema de recomendação aprende e reconhece padrões a partir de dados de treino *offline* do algoritmo (Taneja e Arora 2018; Boussaadi et al. 2020 - 2020). Este método baseia-se em técnicas de *machine learning* como modelos probabilísticos, *lustering*, modelos de fatores latentes e *mining association rules* (Boussaadi et al. 2020 - 2020). Ele é composto por uma fase inicial de aprendizagem, em que são determinados os parâmetros ideais do modelo de recomendação do sistema. Uma vez terminada esta fase, o SR consegue prever as classificações dos itens que o utilizador ainda não tenha classificado.

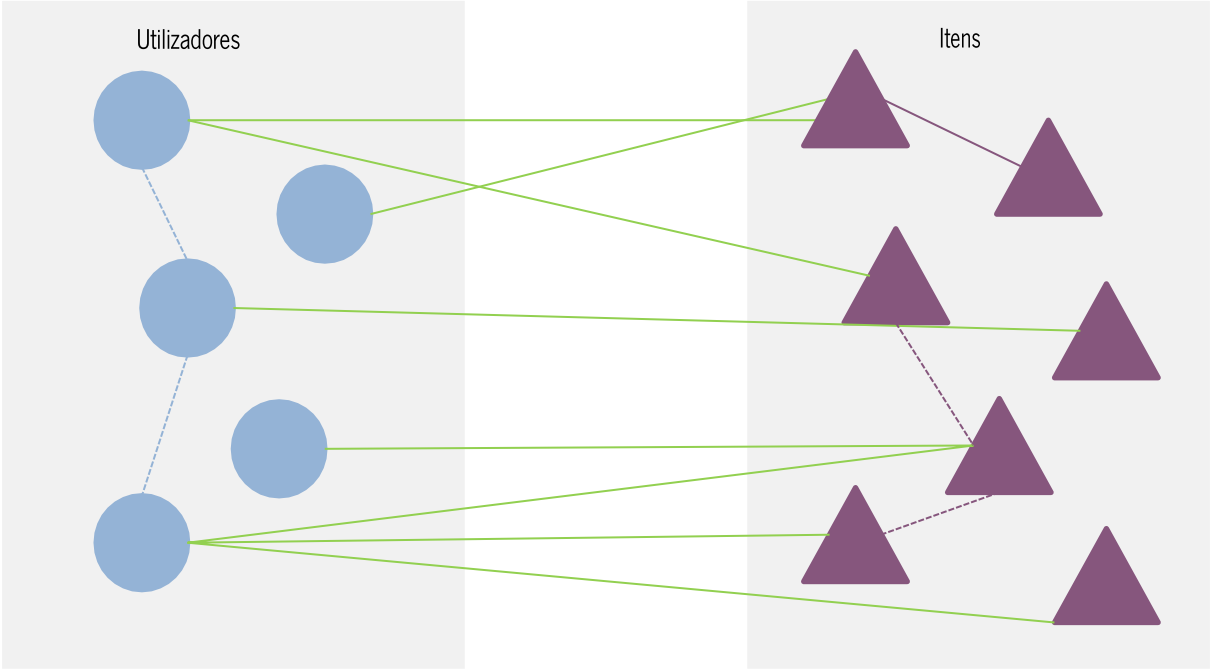
4.2.3 Graph-based

O método *graph-based* (GB) utiliza grafos para representar os dados recolhidos sobre investigadores (utilizadores) e artigos científicos (itens). Aplica, *a posteriori*, um algoritmo de classificação para ordenar os artigos candidatos (Hui et al., 2020).

A recomendação científica GB envolve a representação de um grafo (figura 8) onde os utilizadores e os artigos científicos são arestas (Tang et al., 2021) e as relações de citação entre artigos, por exemplo, são

vértices, ou nós (Ma & Wang, 2019). Na recomendação científica, os grafos podem incluir redes de citação, redes sociais e redes de informação heterogéneas (Tang et al., 2021).

Figura 8 Representação de um modelo de recomendação científica graph-based.



Fonte: Adaptado de Cai et al. (2016).

Este método não só não considera o conteúdo dos ROs mas também as relações entre os artigos. Considera-se que dois artigos são semelhantes quando têm referências em comum ou quando são citados pelo mesmo artigo (Hui et al., 2020). No cenário de redes de citação, um senão é a escassez de dados de citação em artigos científicos, por exemplo, mais recentes.

Ultimamente, as investigações nesta área concentram-se na modelagem gráfica baseada em meta-padrões e na representação de grafos heterogéneos (Ma & Wang, 2019).

Nos grafos baseados em meta-padrões, como o nome indica, são empregados meta-padrões ou meta-grafos de várias semânticas em redes heterogéneas. Contudo, esta abordagem ainda sofre de ineficiência devido às estratégias de extração dos meta-padrões.

Os grafos heterogéneos têm recebido maior atenção talvez porque, como Ma & Wang (2019) notam, as verdadeiras redes de informação científica são, geralmente, grafos heterogéneos, que contêm vários tipos de entidades (p.e. autor, papel, local, tópico) e relações (p.e., de escrita, publicação, colaboração).

É ainda de mencionar as abordagens baseadas em *knowledge graphs*. Um dos seus benefícios é o alívio do problema de escassez de dados graças à tecnologia de *machine learning*, que permite que informações laterais sejam incorporadas nos SRs (Tang et al., 2021). A aprendizagem da representação

de grafos foi introduzida pela rede neural *DeepWalk*, concebida para mapear o contexto das palavras num texto para uma rede (Ma & Wang, 2019).

Tang et al. (2021) dividem os métodos de recomendação de *knowledge graphs* em duas categorias: (1) métodos de conceção e utilização de *semantic paths* e (2) métodos com aprendizagem de representação de caraterísticas. Por exemplo, os métodos (1) podem valer-se de *meta-paths* para calcular a relevância de utilizadores e itens sob *paths* semelhantes. Os métodos (2) mapeiam entidades e relações em vetores de, em simultâneo, baixa dimensão e elevada densidade, de forma a melhorar a representação de utilizadores e itens. Todavia, um defeito destes métodos é que eles desconsideram as associações de alta ordem entre a entidade utilizador e a entidade artigo científico.

4.2.4 Hybrid filtering

Com o propósito de melhorar a precisão das recomendações, aumentar a satisfação do utilizador e aliviar os problemas associados a cada uma das técnicas de recomendação já apresentadas, os SRs híbridos combinam duas ou mais técnicas para beneficiar das vantagens de cada (Sakib et al.2020; Wang et al. 2018). Isto é, o conceito passa por combinar técnicas de recomendação para criar sinergias entre elas.

Na investigação de Wang et al. (2018), foi proposta uma abordagem híbrida para um SR científica que integra informação advinda de redes sociais no seu algoritmo. A informação retirada de *tags* sociais foi utilizada na parte da abordagem CBF para construir os perfis dos investigadores e dos artigos. Já a informação sobre os amigos dos investigadores nas redes sociais foi utilizada na abordagem CF para ajudar a conduzir uma factorização de matriz probabilística unificada. Subsequentemente, o produto de ambas as abordagens foi combinado. Os resultados experimentais obtidos com o conjunto de dados da rede social científica CiteULike provaram que, em comparação a outras abordagens, a proposta obteve os melhores resultados de recomendação.

Kanakia et al. (2019) apresentaram uma plataforma de recomendação científica híbrida utilizada pela Microsoft Academic, combinando abordagens baseadas na co-citação e no conteúdo para maximizar a cobertura, a escalabilidade, melhorar o arranque a frio e a satisfação do utilizador. O *user survey* conduzido para avaliar a proposta mostrou uma correlação positiva evidente entre os valores de similaridade calculados pela plataforma de recomendação híbrida e as pontuações do utilizador. Por outro lado, notou-se uma oportunidade para melhorar a satisfação dos utilizadores relativamente às recomendações apresentadas, particularmente para o método CB.

Yadav et al. (2019 - 2019) propõem RecCite, focando-se na escalabilidade e relevância durante o processo de gerar recomendações. A avaliação da solução dos autores revelou resultados positivos no conjunto de dados ArnetMiner. O algoritmo de recomendação beneficia de uma análise das comunidades de utilizadores através de uma rede de citação. Em adição, beneficia igualmente da mistura dos fatores popularidade (dos artigos) e semelhança semântica no algoritmo personalizado PageRank para produzir recomendações relevantes.

Hui et al. (2020) propõem a abordagem de recomendação híbrida AMHG, que se baseia num grafo heterogéneo de citação multinível. Os autores consideram os artigos-alvo para aliviar o problema do arranque a frio sofrido pelos artigos recentemente publicados e da citação zero. Eles utilizam as relações das citações para extrair a relação implícita entre um artigo e as suas referências; e a informação dos metadados para verificar a semelhança do conteúdo dos artigos-candidatos na rede de citação multinível. Introduzem também a influência dos autores como fator de reordenação da lista de recomendação (Hui et al. 2020).

Na abordagem de Sun et al. (2014), o perfil de utilizador é construído a partir de conteúdo semântico e ligações heterogéneas. Para superar as deficiências dos métodos tradicionais baseados em CB e CF, os artigos-candidatos foram classificados de acordo com a classificação de recomendação agregada, obtida a partir da classificação pré-calculada e da classificação de voto social.

Sakib et al. (2021) propõem uma abordagem híbrida que incorpora metadados contextuais públicos (título, palavras-chave, resumo) na abordagem tradicional CBF para encontrar semelhanças baseadas no conteúdo e informação contextual das relações de citação na abordagem CF. Ambas as similaridades são combinadas para construir o modelo híbrido, que não depende do perfil *a priori* do utilizador.

A proposta de Amami et al. (2017) modela o investigador utilizador com base nos tópicos dos artigos que este classificou, tirando partido dos tópicos latentes nas publicações dos investigadores e das relações entre investigadores da mesma área científica. A abordagem híbrida combina a análise de conteúdo baseada em modelação probabilística e técnicas da CF baseadas num modelo linguístico.

Ma & Wang (2019) desenvolveram uma recomendação científica baseada na representação gráfica heterogénea que considera as entidades heterogénicas e as relações dos gráficos académicos, bem como a informação do conteúdo dos artigos científicos na representação do utilizador. O algoritmo primeiro constrói o perfil dos utilizadores e dos artigos científicos com a extração de informação sobre o conteúdo do artigo. A doc2vec é posteriormente utilizada para as representações dos nós das relações nos grafos heterogéneos.

4.3 Problemas e soluções das técnicas de recomendação

4.3.1 *Content-based*

Devido à lógica da CBF, os SRs só podem gerar recomendações úteis quando as características dos itens e as preferências do utilizador estão disponíveis. Caso contrário, verificar-se-á o problema do arranque a frio (Jugovac e Jannach 2017). O problema do arranque a frio ocorre quando há falta de informação sobre:

- um novo utilizador que ainda não interagiu com os itens, e cujo corpus é considerado vazio,
- um novo artigo que não foi objeto de interesse por nenhum utilizador.

Por outro lado, a CBF considera os interesses atuais dos utilizadores. Significa que, se os seus interesses se transfigurarem, as recomendações refletem essa mudança (Boussaadi et al. 2020 - 2020). Descreve-se de seguida outros dois problemas comuns em SRs com CBF. No sentido de aumentar a utilidade do tópico, mencionam-se soluções encontradas ou desenvolvidas por investigadores.

Sobreespecialização e baixa serendipidade

A sobreespecialização e a deficiência de serendipidade são dos desafios mais substanciais dos SRs baseados no conteúdo.

A sobreespecialização consiste na recomendação de itens demasiado similares entre si (Taneja e Arora 2018; Saat et al. 2018). Ainda que resulte isso em itens relevantes, há o risco de as recomendações não serem tão úteis devido à ausência de novidade e imprevisibilidade, uma vez que elas estão limitadas aos itens previamente classificados pelo utilizador. A solução passa por incluir serendipidade, ou seja, novidade, imprevisibilidade e relevância no SR (Chen 2004).

A serendipidade pode ser interpretada quase como o inverso da similaridade dado que a sua introdução no SR resulta em recomendações de itens novos, com os quais o utilizador não está familiarizado.

Os SRs que utilizam abordagens de *Linked Open Data* (LOD) evidenciam ser boas possibilidades para encontrar itens relevantes, inesperados e novos num grande conjunto de dados (Saat et al. 2018). Em Maccatrozzo et al. (03/07/2017), os autores propõem o modelo SIRUP, focado na serendipidade em SRs baseados no conteúdo. A novidade dos itens é calculada com o *cos*, utilizando caminhos de LOD. A avaliação, aplicada a programas televisivos, mostram que o SIRUP permite identificar recomendações serendipistas e, em simultâneo, ter 71% de precisão. Nota-se, no entanto, uma lacuna no que concerne o estudo do impacto de LOD nos SRs com CBF, inclusive na área científica.

Semântica

Normalmente, as caraterísticas extraídas a partir de metadados, ou caraterísticas textuais, são os atributos que descrevem um item (Beel et al. 2016) — ainda que as caraterísticas possam ser também não-textuais, como é exemplo o layout da informação. O que se verifica nestes casos é a extração de conteúdo pouco significativo para corretamente capturar a semântica dos interesses do utilizador. Apesar de os sistemas de CBF serem concebidos, principalmente, para recomendar itens baseados em texto, eles criam complicações que têm origem na ambiguidade da linguagem, tais como polissemia e sinonímia (Son e Kim 2018).

Noutra perspetiva, utilizar somente o título e resumo do artigo candidato pode conduzir a recomendações irrelevantes. Nem sempre os títulos e/ou os resumos dos artigos refletem adequadamente as contribuições da investigação desenvolvida. Uma solução para este problema é ponderar o conteúdo completo do item candidato a recomendação, tal como o método proposto por Hanyurwimfura et al. (2015), que considera o conteúdo integral dos artigos para gerar recomendações. Os autores propõem um SR académico sem necessidade de perfis de utilizador. A abordagem toma um único artigo científico como input, do qual se extraem os tópicos principais e as frases-chave. Estas duas caraterísticas são posteriormente submetidas como termos/frases de pesquisa a bases de dados e repositórios online para reunir os artigos semelhantes (Hanyurwimfura et al. 2015).

Ferrara F., Pudota N., Tasso C. (2011) apresentaram um SR científica que produz perfis de utilizador adaptáveis bem como descrições semânticas de itens, através da extração das frases-chave dos artigos científicos. Os autores pressupõem que as frases-chave possuem informação contextual significativa o suficiente para melhorar o mecanismo de filtragem. As descrições dos itens são comparadas pelo cos para avaliar a relevância de um novo documento no que diz respeito aos interesses do utilizador.

4.3.2 Collaborative-filtering

Descreve-se de seguida 4 problemas comuns aos SRs de *collaborative filtering*, nomeadamente o problema de arranque a frio, que corresponde à falta de informação sobre os utilizadores ou itens para gerar recomendações, de escassez de dados e de adaptação à mudança de interesses por parte dos utilizadores. Para cada um deles são exploradas soluções propostas por autores, no contexto de recomendações científicas.

Arranque a frio: novos utilizadores

Como notado por Sakib et al. (2020), a maioria dos SRs dependem de perfis de utilizador *a priori* – perfis de utilizador criados com base no histórico – e, por consequência, não conseguem recomendar itens a novos utilizadores. Em acréscimo, como a maioria deles também utiliza informação contextual não-pública, as restrições dos direitos de autor impedem de encontrar semelhanças adequadas entre papeis. Procurando colmatar esses problemas, os autores desenharam um SR científica que utiliza a filtragem colaborativa, não depende de perfis de utilizador *a priori*, e utiliza somente informação contextual pública. As associações entre os itens foram encontradas através de relações de 2 níveis de papel(item)-citação. Outra solução é a de Haruna et al. (2017), exposta na secção *Adaptação aos interesses em mudança dos utilizadores* em maior detalhe, que não requer um perfil de utilizador *a priori* graças à utilização de metadados contextuais disponíveis publicamente. Resumidamente, os autores introduzem uma abordagem colaborativa que deduz associações ocultas entre um item-alvo e as suas referências, bem como entre as citações do item-alvo utilizadores.

Arranque a frio: novos itens e falta de classificações

A CF não pode gerar recomendações precisas sem classificações suficientes dos itens (Son e Kim 2018; Chen 2004), algo que acontece com frequência no começo da implementação dos sistemas de recomendação devido à típica reduzida adesão inicial.

Sem classificações, o sistema não pode gerar recomendações, e sem ou com poucos utilizadores, não há forma de introduzir novas classificações no sistema. Trata-se, de facto, de um problema interligado com o arranque a frio visto anteriormente. Ademais, se há uma insuficiência de classificações e, consequentemente, incapacidade de recomendação, os utilizadores sentem relutância e apreensão em continuar a utilizar o SR.

Para resolver o problema, Kautz et al. (1997) criaram uma matriz de classificações a partir da rede de citações entre artigos científicos. Nesta matriz, cada linha corresponde a um autor, cada coluna corresponde a um artigo, e cada célula corresponde a uma classificação. Haruna et al. (2017) extraíram a pontuação de classificação entre investigadores e trabalhos de investigação com base nas relações item-citação. Adicionalmente, as associações entre artigos científicos podem ser depreendidas através das classificações explícitas dos utilizadores, mas também de ações implícitas, da análise de citações ou de citações. McNee et al. (2002) testaram a abordagem das classificações com origem em citações. Nesse caso, na sua matriz, um artigo representa um utilizador e uma citação representa um item. A

pontuação é definida pelas citações encontradas na lista de referências do artigo. Ao se ter os utilizadores como "artigos", garante-se que cada utilizador fornece classificações (McNee et al. 2002).

As ações implícitas ocorrem nas interações entre utilizadores e itens. Yang et al. (2009) focaram-se no número de páginas lidas pelos utilizadores, seguindo o raciocínio: quantas mais páginas lidas, maior o interesse, maior a relevância e melhor a classificação. Em contraste, Pennock et al. (2000) equacionaram o ato de download, a adição de um artigo ao perfil de utilizador, a edição de detalhes dos artigos científicos, e a visualização da bibliografia a avaliações positiva. McNee et al. (2002) postularam que quando dois autores citam os mesmos artigos, ambos são semelhantes na sua lógica. Ou, se um utilizador lê ou cita um artigo, as citações do artigo citado vão de encontro aos gostos e interesses do utilizador. Contudo, nada garante que qualquer uma destas suposições em relação às ações implícitas dos utilizadores corresponda à realidade, o que pode acabar por ter um efeito negativo na qualidade das recomendações.

Adaptação aos interesses em mudança dos utilizadores

Os sistemas de recomendação de CF baseiam-se no pressuposto de que os utilizadores com gostos semelhantes no passado terão os mesmos gostos no futuro. No entanto, é natural que os investigadores publiquem *research outputs* em mais do que uma área científica ou que alterarem as suas áreas de interesse ao longo da carreira. Pode-se, assim, tornar difícil encontrar um grupo de utilizadores semelhante e recomendar artigos adequados ao utilizador-alvo quando uma base de dados contém muitos desses casos (Son e Kim 2018).

Haruna et al. (2017) apresentam uma abordagem colaborativa para um SR científica que deduz associações ocultas entre um item-alvo e as suas referências, bem como entre os artigos que citam o item-alvo, a fim de personalizar as recomendações. Este ato é possível através de metadados contextuais disponíveis publicamente. A proposta admite a lógica de que, se dois artigos científicos pertencem às referências bibliográficas de um número significativo de artigos em comum, então eles são semelhantes. Na abordagem proposta, um item-candidato é qualificado para consideração se e só se citar qualquer uma das referências do item- alvo e se existir outro item que cite simultaneamente o candidato e o item-alvo. São depois recomendamos os top-N itens mais semelhantes ao item-alvo. Significa isto que o a abordagem proposta é capaz de gerar recomendações personalizadas, independentemente das áreas científicas de interesse do utilizador e da sua experiência profissional.

Escassez de dados

O desempenho dos modelos de CF diminui consideravelmente quando o SR não tem dados de feedback suficientes sobre utilizadores ou itens. Tenha-se como exemplo os utilizadores que dão feedback explícito ao SR quando avaliam os itens numa escala de 0 a 10 estrelas. É a partir destes dados que se cria o perfil do utilizador e se geram recomendações. Contudo, depara-se com um problema quando ele não fornece esse feedback necessário, observando-se uma escassez de dados (tabela 10) ou *data sparsity* (Alfarhood e Cheng 2019 - 2019; Amami et al., 2017; Son e Kim 2018). Quanto maior a escassez, mais inexatas as previsões.

Tabela 10 Exemplificação de escassez de dados.

Id utilizador	Id item				
	2115	248586	8563	852	655220
131	NaN	NaN	NaN	NaN	NaN
251	NaN	NaN	NaN	3	NaN
3841	NaN	4	NaN	NaN	NaN

Para solucionar este problema, Alfarhood e Cheng (2019 - 2019) propuseram a técnica de recomendação híbrida *Collaborative Attentive Autoencoder* (CATA). Ela aproveita a informação textual dos itens para aprender os seus fatores latentes através de um mecanismo de atenção – uma técnica de *deep learning* – que captura a parte mais pertinente da informação. Permite, assim, prever com mais precisão as classificações dos utilizadores e gerar melhores recomendações. Fatores latentes representam categorias presentes nos dados. Num *dataset* de artigos científicos, "biomecânica", "geologia" e "história da arte" podem ser fatores latentes, que representam um domínio científico. A avaliação realizada ao CATA permitiu concluir que o desempenho deste aumenta consistentemente à medida que a escassez de dados também aumenta.

4.3.3 Graph-based

Esta secção discorre os problemas mais predominantes dos SRs *graph-based*: o arranque a frio, a dificuldade em modelar associações de alta ordem entre as entidades, escassez de dados e as estratégias ineficientes de extração de meta-padrões.

Arranque a frio

A proposta de Wang et al. (2020) parte da ideia de que o ato de estabelecer correlações entre documentos se deve basear em atributos como o título, autor, resumo, palavras-chave, relações de citação referências, e de que os *knowledge graphs* podem resultar não só em resultados mais precisos, mas podem também aliviar a escassez de dados. Os autores constroem um *knowledge graph* de artigos científicos que modela as preferências do utilizador de acordo com a interação histórica deste com os documentos. O modelo neuronal *long short-term memory* é utilizado para extrair informação entre os nós e combinar com as preferências de utilizador para obter a lista de recomendação.

Cai et al. (2016) apresentam o modelo grafo bi-relacional UAGMT para resolver o problema do arranque a frio de novos itens. Para colmatar a falta de informação histórica dos novos artigos científicos, o UAGMT integra informação de *tags*. Primeiro, o modelo utiliza o título e o resumo do artigo para sumariar o seu conteúdo, adotando a premissa de que a informação das *tags* pode ser combinada com outro tipo de informação para calcular a similaridade entre artigos e gerar recomendações. Na representação gráfica, existem ligações *intra-layers* e ligações *inter-layers*. No primeiro caso, as ligações denotam as relações utilizador-artigo e, no segundo caso, as ligações denotam as relações entre o mesmo tipo de objetos. Por último, para recomendar novos artigos, é aplicado o processo matemático passeio aleatório (*random walk*) em combinação com um algoritmo de reinício.

Associações de alta ordem entre as entidades e escassez de dados

O método de Tang et al. (2021) resolve dois aspetos: (1) recomendar artigos relevantes sabendo que o histórico de interação entre SR e utilizador é escasso, (2) recomendar artigos relacionados com o domínio científico de investigadores com conhecimentos limitados sobre o domínio. Os autores propõem o método *Content-based and knowledge Graph-based Paper Recommendation method* (CGPRec), que utiliza a técnica de processamento em linguagem natural *Word2vec* e uma *Double Convolutional Neural Network* para processar texto, e um módulo de extração de características com duas camadas de auto-atenção para obter as preferências explícitas do utilizador. Com uma base de dados de conhecimento externa, o CGPRec extrai conceitos do conteúdo dos artigos científicos para construir um *knowledge graph*. Esse grafo é composto por nós de utilizador, nós de artigos científicos e nós de metadados. Tang et al. (2021) propõem ainda uma rede convolucional para modelar associações de alta ordem no *knowledge graph* e assim captar as preferências implícitas do utilizador.

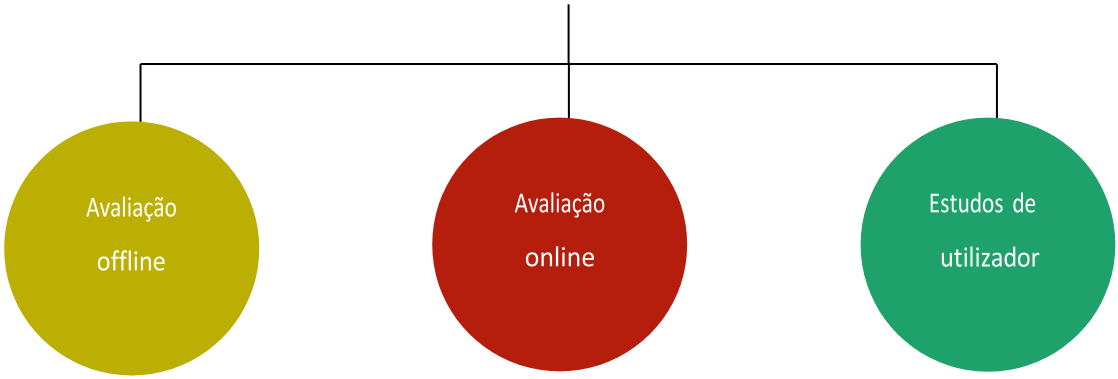
Ineficiência das estratégias de extração para meta-padrões

Ma & Wang (2019) resolvem o problema de recomendações científicas personalizadas, no contexto de redes de informação heterogéneas, com o HGRec. O método baseia-se na aprendizagem de representação gráfica heterogénea. Primeiro, o perfil dos utilizadores e dos artigos científicos são construídos a partir da extração de informação do conteúdo dos artigos. De seguida, aplicam a ferramenta de programação neurolinguística Doc2vec para representar os nós do utilizador e do artigo. Por fim, as recomendações resultam do cálculo da semelhança do *cos* entre os vetores das caraterísticas do utilizador e os vetores das caraterísticas do artigo científico.

4.4 Avaliação dos sistemas de recomendação

Nesta secção são descritos três tipos de avaliação (figura 9) que se podem aplicar aos sistemas de recomendação para perceber o estado da performance do sistema: 1) avaliação offline, 2) estudos de utilizador e 3) avaliação online. Em adição, são discutidos os recursos necessários e procedimentos necessários para realizar cada um destes tipos de avaliação.

Figura 9 Tipos de avaliação aos sistemas de recomendação.



4.4.1 Avaliação offline

A avaliação offline permite medir a qualidade das recomendações geradas pelo SR a custos reduzidos, em especial, quando comparada com a avaliação offline e estudos de utilizador. Ela não exige que o sistema esteja implementado nem que sejam utilizados utilizadores reais, razões pelas quais a avaliação offline é amplamente aplicada. Contudo, por essas mesmas razões, (1) só um número reduzido de perguntas, normalmente relacionadas com a capacidade de previsão de um algoritmo, pode ser

respondido e (2) a influência do sistema de recomendação no comportamento do utilizador não pode ser diretamente medida.

Devido ao acesso limitado a dados do mundo real, os SRs são maioritariamente avaliados offline através de *datasets*. Os *datasets* para recomendação científica contêm dados sobre autores e artigos, por exemplo, id de autor, interesses de investigação do autor, afiliações, título, classificações, ano de publicação, jornal, citações e URLs de PDFs públicos.

Visto o objetivo da avaliação offline ser simular o comportamento dos utilizadores que interagem com o SR, o *dataset* deve permitir a comparação das classificações previstas pelo sistema com as classificações reais (Barros et al. 2019). Desta forma, os dados do *dataset* devem ser o mais compatíveis possível com os dados com que o sistema lidará na vida real. Evita isto o enviesamento na distribuição de utilizadores, itens e classificações (Ricci et al. 2015).

A tabela 11 referencia *datasets* científicos que foram identificados durante o processo da revisão de literatura, assim como a URL onde eles podem ser descarregados. Eles encontram-se disponíveis online até ao presente momento, pelo que podem ser utilizados para realizar uma avaliação offline a um SR científica.

Tabela 11 Datasets *para sistemas de recomendação científica.*

DATASET	**URL**
Anthology Reference Corpus	https://www.aclweb.org/anthology/ https://drive.google.com/file/d/0B2Mzhc7popBgTjRRX1hPY2g2MUE/edit
ACM portal	https://dl.acm.org/
ArnetMiner	https://www.aminer.org/aminernetwork
DBLP	https://dblp.uni-trier.de/xml/
CITESEERX	https://csxstatic.ist.psu.edu/downloads/data
Docear	https://docear.org/labs/
RARD	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HA 8EAH
RARD II	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AT4 MNE
Sugiyama e Kan (2013)	https://scholarbank.nus.edu.sg/handle/10635/146027
unarXive	https://zenodo.org/record/3385851

Na avaliação offline, o *dataset* é dividido em conjunto de dados para treino (*training set*), destinado a treinar os algoritmos do SR, e conjunto de dados para teste (*test set*) (Sakib et al. 2020).

O conjunto de dados para teste serve como ponto de referência para aquilo que é aceitável, ou seja, ele determina a importância dos resultados produzidos pelo conjunto de treino ou o desempenho do algoritmo de recomendação. O conjunto de dados para teste contém classificações reais que são consideradas como valor base. Idealmente, a avaliação offline demonstra que as classificações previstas com o conjunto de dados para treino (tabela 12) coincidem com a verdade/realidade. Isto implica que as interações do utilizador com o sistema sejam ocultadas (tabela 13). Assim sendo, a verdade/realidade corresponde às classificações ocultadas no conjunto de dados para teste.

Tabela 12 Matriz de classificação do conjunto de dados para treino.

	Item1	Item2	Item3	Item4
Utilizador1	-1	+1	+1	
Utilizador2		+1	-1	+1
Utilizador3	+1	-1	-1	
Utilizador4	+1	-1	+1	+1

Tabela 13 Matriz de classificação do conjunto de dados para teste.

	Item1	Item2	Item3	Item4
Utilizador1	-1	+1	+1	
Utilizador2		+1	-1	+1
Utilizador3	+1	-1	-1	
Utilizador4	+1	-1	+1	?

As classificações no conjunto para teste são inputs no SR para a construção de uma matriz de interação entre o utilizador e os itens. É através delas que o sistema atribui uma pontuação a cada um dos itens presentes no conjunto de dados de teste associado a um utilizador específico, ou a cada um dos itens ausentes no conjunto de dados de treino associado a um utilizador específico. Frisando novamente a ideia, com o conjunto de dados para treino, o SR tenta prever as classificações no conjunto de dados para teste. As classificações previstas são depois comparadas com as classificações reais no conjunto de teste. Para analisar o desempenho do algoritmo de recomendação, diferentes métricas podem ser calculadas

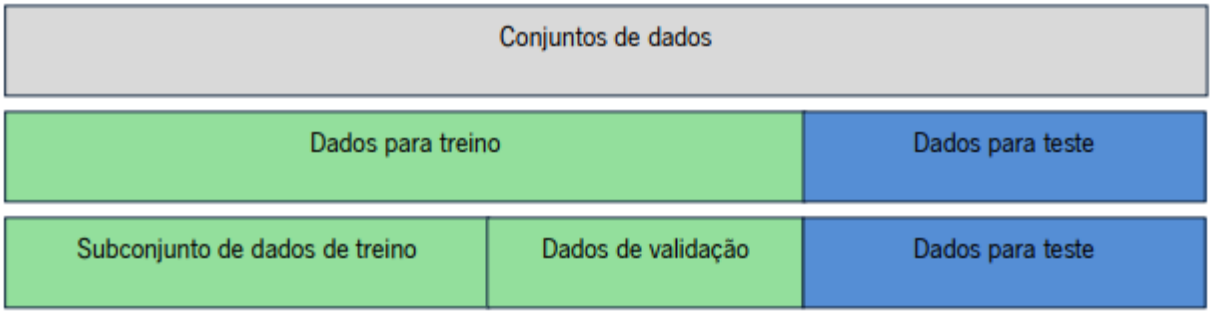
Datasets e divisão dos datasets

O primeiro passo para a avaliação offline é escolher um *dataset*. É de notar que, à medida que o volume de dados aumenta, maior é a dificuldade dos algoritmos em lidar com a escassez de dados, arranque a frio, escalabilidade, diversidade, entre outros.

O *dataset* selecionado é dividido em dados para treino e teste do SR. Por sua vez, os dados para treino descortinam-se, normalmente, num subconjunto de dados de treino e dados de validação (figura 10). Idealmente, o SR é avaliado com amostras de dados que não foram aproveitados para treinar ou melhorar o seu modelo, de modo a evitar o enviesamento. Os dados de validação servem esse propósito, qualificando o desempenho dos SRs imparcialmente. Em contrapartida, os dados de treino treinam o modelo.

Importa salientar que, na literatura, o termo "conjunto de validação" é frequentemente empregue enquanto sinónimo de "conjunto de teste", embora uma distinção pareça, do ponto de vista pessoal, mais benéfico para a compreensão do processo aqui descrito.

Figura 10 Representação visual do processo de divisão dos dados para avaliação.



A divisão do conjunto de dados pode ser feita de acordo com uma das seguintes estratégias (Jiang et al. 2012; Wang et al. 2006; Wang et al. 2018; Sun et al. 2014; Sakib et al. 2020; Chen 2004):

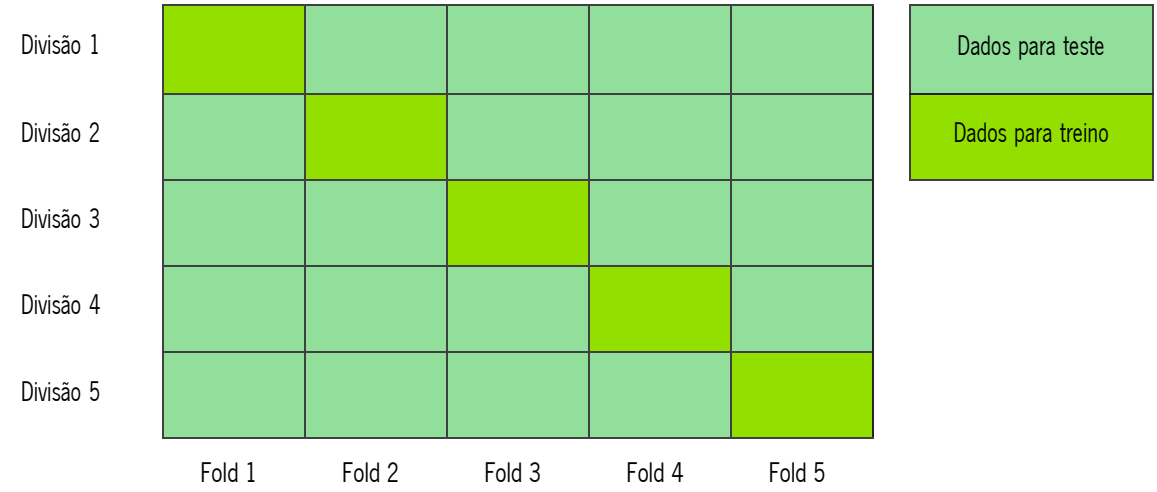
- ***Leave-one-last:*** extrai a interação final do utilizador como conjunto de dados para teste, onde a penúltima interação é normalmente utilizada como validação e as restantes transações são utilizadas para treino do SR.
- **Divisão aleatória**: implica selecionar de forma aleatória o limite de dados para treino/teste por utilizador. Por exemplo, Wang et al. (2006) recolheram amostras aleatórias dos dados para limitar o número de utilizadores a 428 e o número de itens a 516.
- **Divisão dos utilizadores:** consiste em dividir o conjunto de dados por utilizador. Um conjunto de utilizadores (e respetivos atributos) é reservado para treino do SR, enquanto que um outro conjunto de utilizadores é utilizado para testes. Requer que o sistema não sofra de arranque a frio.
- **Divisão baseada no tempo**: considera uma linha temporal de acontecimentos definida, a título de exemplo, pelas interações entre utilizador e item. Um possível procedimento é definir um ponto (fixo) no tempo e fragmentar todas as interações prévias a esse ponto (dados para

treino) das interações posteriores a esse ponto de tempo (dados para teste). Existem duas variações para definir esse ponto de separação:

  - Divisão específica ao utilizador: assegura o mesmo rácio por utilizador com a seleção de uma percentagem das últimas interações de cada utilizador (ex.: 20% das últimas interações de cada utilizador é reservada para testes)
  - Divisão global: onde ele é definido por um ponto de tempo fixo partilhado por todos os utilizadores, de modo a se obter um rácio de treino/teste global (ex.: divisão do conjunto de dados em 80%/20%).

- **Validação cruzada k-fold** (figura 11): tem um único parâmetro k, sendo k o número de subconjuntos em que um conjunto de dados se divide. Uma vez feita essa divisão, cada subconjunto pode ser utilizado como conjunto de dados para teste enquanto todos os outros subconjuntos juntos são utilizados como conjunto de dados para treino. Por exemplo, Sakib et al. (2020) dividiram o seu conjunto de dados em 5 grupos, selecionando 20% da amostra como um conjunto de teste (equivale a 0,20*5=1). Pode-se assim dizer que Sakib et al. (2020) executaram uma validação cruzada com k=5, ou ainda, validação cruzada de *5-fold*, e que 1 dos 5 subconjuntos foi utilizado para teste. Note-se o nome desta estratégia. Ele relembra os dados para validação e parece esquecer o treino do modelo do SR. Na verdade, a estratégia de validação *k-fold* é uma forma de avaliar o SR sem enviesamento que não requer a divisão dos dados para treino em dados de treino e dados de validação.

Figura 11 Conceito de validação cruzada k fold, para k=5.



Esta lista não é exaustiva, pelo que existem outras estratégias que podem ser tidas em consideração na altura de divisão dos *datasets*, nomeadamente estratégias de validação cruzada, que se subdividem para além do *k-fold*.

Métricas de avaliação offline

Numa avaliação offline típica são selecionados conjuntos de dados para treino e para teste a partir de um *dataset*. Ocorrida a etapa, é construído um modelo preditivo nas amostras de treino, que é posteriormente avaliado com amostras de validação.

O desempenho geral do SR em escrutínio é calculado através de diferentes métricas de avaliação. Inferir acerca o SR a partir de um só indicador é insuficiente, pelo que se recomenda perspetivar segundo diferentes métricas de:

- precisão preditiva,
- precisão de classificação (ou de apoio à decisão) e
- precisão top-N (ou de rank).

Alguns algoritmos procuram prever a classificação dos itens a partir do feedback explícito do utilizador. Caso o SR não sofra de escassez de dados, métricas de avaliação como o *mean absolute error* (MAE) ou o *root mean squared error* (RMSE) são adequadas (Barros et al. 2019). Métricas de precisão preditiva como estas duas medem o desempenho do algoritmo de recomendação comparando a previsão do algoritmo com a classificação real de um item, dada por um utilizador.

Se o erro absoluto (*absolute error* - AE) é o número de erros observado nas medições, ou seja, a diferença entre o valor previsto computado e o valor verdadeiro, o MAE (equação 13), aplicado aos SRs científica, é o desvio médio absoluto entre uma classificação prevista e a classificação real de um item dada por um utilizador. Quanto mais baixo o valor, melhor o algoritmo. Em termos matemáticos, o MAE é a soma do valor absoluto das diferenças entre os valores esperados e os valores previstos, dividido pelo número total de previsões. A equação 13 traduz esta definição em termos matemáticos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y| \qquad (11)$$

onde n representa o número de erros, y_i a classificação prevista, y a classificação real, e $|yi - y|$ os erros absolutos.

O RMSE (equação 14) é o desvio padrão dos erros de previsão, ou resíduos. É uma métrica de precisão preditiva que indica quão longe os as classificações previstas estão da linha de regressão. O RMSE informa como os dados estão concentrados em torna da linha de melhor ajuste (classificações reais). A fórmula correspondente é dada pela equação 14, onde N denota o tamanho da amostra, y_i' as classificações previstas, y_i' os valores reais, e $(y_i' - y_i)^2$ as diferenças entre as classificações previstas e reais, ao quadrado.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i' - y_i)^2}{N}} \quad (12)$$

O MAE e o RMSE podem ser utilizados em conjunto para aferir a variação dos erros num conjunto de previsões. De ter em mente que o resultado do RMSE é sempre maior ou igual ao MAE. No caso do MAE, os valores de erro individuais seguem um comportamento linear – um erro de 10 contribui 2 vezes mais do que um erro de 5. Por sua vez, o RMSE torna os valores entre 0 e 1 ainda mais pequenos e os valores maiores, maiores, ou seja, os valores de erro elevados são ampliados, enquanto que os valores pequenos são ignorados.

As métricas de precisão preditivas preocupam-se com o quão corretamente o algoritmo do SR consegue prever as classificações dos utilizadores. As métricas de precisão de classificação, também conhecidas por métricas de apoio à decisão, reconhecem que, para muitos SRs, o objetivo é ajudar os utilizadores a tomar boas decisões. Estas métricas, por exemplo, precisão, *recall* e *F-measure* (F1), focam-se em distinguir se as classificações previstas são boas ou más, na ótica do utilizador.

Os valores da precisão, *recall*, e F1 variam entre 0 e 1. O algoritmo melhor é quanto mais próximo de 1. Para um determinado número k de itens recomendados, a precisão (equação 15) é a proporção de itens recomendados relevantes para o utilizador (Hui et al. 2020). A precisão em k, P@k, é a fração dos itens top-k obtidos relevantes para um utilizador.

$$Precisão = \frac{Número\ de\ itens\ relevantes\ recomendados}{Número\ total\ de\ itens\ recomendados} \quad (13)$$

O *recall* (equação 16) é a proporção do número total de itens relevantes recomendados no número total de artigos na base de dados (Sakib et al. 2021). Por exemplo, se uma lista de tamanho 10 recomenda 5 itens relevantes para o utilizador j, cujo número total de itens relevantes nesse conjunto de teste é 5, então, o *recall* é de 100%, uma vez que o algoritmo recomenda todos os itens possíveis em que o utilizador teve interesse.

O *recall* é, grande parte das vezes, utilizado para o feedback implícito, todavia a precisão não é indicada para tal pois assume que o valor 0 na matriz utilizador-item significa que o utilizador não está interessado no item quando pode também significar que ele ainda não conhece o item (Alfarhood e Cheng 2019 - 2019).

$$Recall = \frac{Número\ de\ itens\ relevantes}{Número\ total\ de\ itens\ relevantes} \quad (14)$$

O F1 (equação 17) é a média harmónica da precisão e *recall*, o que permite uma avaliação global do algoritmo de recomendação.

$$F1 = \frac{2 \times Precisão \times Recall}{Precisão + Recall} \quad (15)$$

Alguns SRs apresentam aos seus utilizadores recomendações na forma de listas compostas por N itens. Métricas de top-N como o *normalized discounted cumulative gain* (nDCG), *mean average precision* (MAP), e *mean reciprocal rank* (MRR) visam capturar a qualidade de uma classificação em particular, tendo em conta a preferência expressa pelo o utilizador por determinados itens – geralmente os itens presentes no conjunto de dados para teste e classificados acima de um limiar pré-determinado.

Ao contrário das métricas já vistas, o MAP, o MRR e o nDCG consideram o *ranking* dos itens relevantes na lista de recomendação (Sun et al. 2014). Fornecem, portanto, uma visão da capacidade do sistema em devolver um item relevante no topo do ranking.

Se a *average precision* (AP) é a média dos valores de precisão em todos os *ranks* onde se encontram itens relevantes, o MAP é a média de todos os APs. O MAP corresponde à equação 18, onde k é o número de itens recomendados; N é o tamanho da lista de recomendação; U denota o número total de utilizadores registados no SR; m_j é o número de itens relevantes ao utilizador j na lista de recomendação; e $P(R_{jk})$ representa a precisão dos resultados obtidos a partir do resultado superior (o primeiro) até chegar ao item k.

$$MAP = \frac{1}{U}\sum_{i=1}^{U}\frac{1}{m_j}\sum_{k=1}^{N}P(R_{jk}) \quad (16)$$

O MRR indica onde, no ranking, o primeiro item relevante é devolvido pelo SR, com uma média sobre todos os utilizadores (Hui et al. 2020; Sakib et al.2021). A fórmula da métrica pode ser consultada na equação 19, onde U denota o número total de utilizadores registados no SR; j representa os utilizadores do sistema; e F_j é a posição do primeiro item relevante para o utilizador j.

$$MRR = \frac{1}{U}\sum_{j=1}^{U}\frac{1}{rank\ F_j} \quad (17)$$

O nDCG é uma medida que dá mais importância a itens altamente classificados e incorpora diferentes níveis de relevância (Sugiyama e Kan 2010). Pode-se afirmar que ele avalia a qualidade do ranking, sendo que os itens com classificação mais elevada devem aparecer em primeiro lugar no ranking.

Imagine-se que o SR retorna 3 recomendações, que são classificadas pelo utilizador numa escala de 0 a 2, em que 0 corresponde a irrelevante, 1 corresponde a algo relevante e 2 corresponde a extremamente relevante. Para a fórmula nDCG, primeiro é calculado o *cumulative gain* (CG) de forma a obter a soma

das classificações. Para tal utiliza-se a equação 20, onde i é o item recomendado, e rel_i é a relevância do item recomendado, consoante a classificação dada pelo utilizador a i.

$$CG = \sum_{i=1}^{3} rel_i \tag{18}$$

Com o resultado obtido calcula-se o *discounted cumulative gain* (DCG) através da equação 21. Em DCG, a soma segue a ordem das recomendações geradas, ou seja, soma-se o DCG da primeira recomendação (do top N) com o DCG da segunda recomendação, com o DCG da terceira recomendação.

$$DCG = \sum_{i=1}^{3} \frac{rel_i}{\log_2(i+1)} \tag{19}$$

Os itens precisam agora de ser reordenados por ordem decrescente de acordo com a classificação (de 1 a 2) que obtiveram, para se calcular o *Ideal Discounted Cumulative Gain*

$$IDCG = \sum_{i=1}^{3} \frac{(2^{rel_i} - 1)}{\log_2(i+1)} \tag{20}$$

Por último, calcula-se o *Discounted Cumulative Gain* normalizado (equação 23).

$$nDCG = \frac{DCG_3}{IDCG_3} \tag{21}$$

Segundo Yadav et al. (2019 - 2019), no contexto de recomendações científicas, a precisão e a escalabilidade são as métricas mais relevantes. A escalabilidade representa a capacidade do SR em trabalhar eficientemente num ambiente com numerosos utilizadores e itens. Ela é, geralmente, calculada como o tempo médio de resposta, ou seja, o tempo que o sistema demora para responder ao pedido do utilizador, neste caso a geração de recomendações.

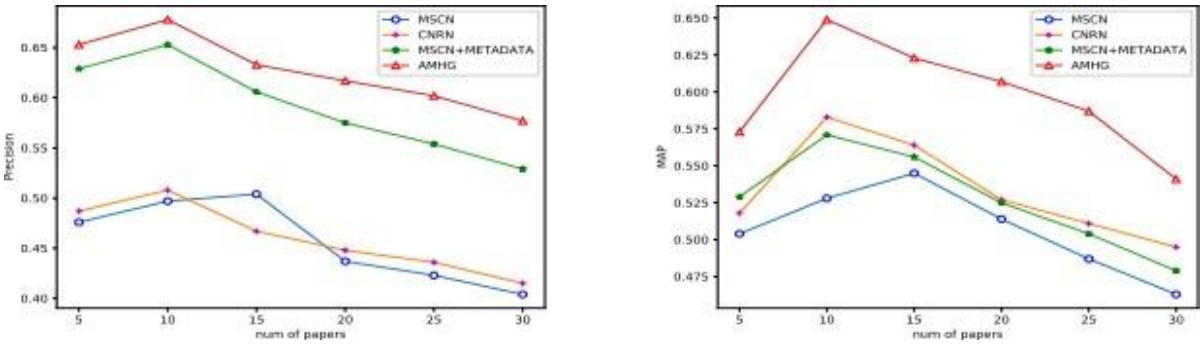
Seleção de baselines para avaliar o SR

Os resultados das métricas, por si só, carecem de significado. Eles unicamente ganham sentido na presença de um limite numérico indicativo da qualidade do desempenho do SR. Tal limite é definido pelos resultados das métricas de outros algoritmos-candidatos ou de algoritmos do estado da arte ou comparáveis com o algoritmo proposto.

Quando se comparam dois ou mais algoritmos candidatos, prevalece aquele que tiver resultados de maior magnitude nas métricas. Quando se avalia uma abordagem, em especial, inovadora, deve-se efetuar a comparação com uma referência-base *(baseline)* representativa das abordagens do estado-da-arte. Deste modo, é possível quantificar o quão mais eficaz uma abordagem é. A figura 12 mostra um

exemplo de comparação dos resultados da abordagem proposta por Hui et al. (2020) com os resultados de *baselines*, através da métrica precisão (à esquerda) e da métrica MAP (à direita).

Figura 12 Exemplo da comparação de uma proposta de abordagem com referências-base.



4.4.2 Avaliação online

A avaliação online é uma avaliação centrada no utilizador. As métricas da avaliação offline são uma tentativa de medir a experiência do utilizador. Porém elas não exprimem todo o contexto de uso nem o papel das preferências e comportamento do utilizador na sua interação com o SR.

O tipo de avaliação que proporciona as provas mais convincentes quanto ao verdadeiro valor do SR é a avaliação online, onde o sistema é utilizado por utilizadores reais que executam tarefas reais, tipicamente inconscientes da sua participação na experiência (Ricci et al. 2015).

O aspeto mais positivo desta abordagem é ela imitar a vida real. Todavia, este tipo de avaliação exige que se tenha acesso a um SR já desenvolvido e disponível publicamente, para além de que estas avaliações são mais morosas e dispendiosas (Hui et al. 2020).

As avaliações online medem as taxas de aceitação de recomendações nos SRs do mundo real (Hui et al. 2020). A taxa de aceitação é normalmente calculada através de feedback implícito sobre a satisfação do utilizador. A questão do feedback implícito já foi discutida em pormenor ao longo da revisão de literatura, mas o pressuposto é que se um utilizador clica, descarrega, ou compra um item recomendado, é porque ele gosta da recomendação. Esta lógica tem os seus defeitos: o utilizador pode comprar um artigo (item recomendado) e manifestar uma opinião negativa após a sua leitura, ou o utilizador pode clicar num artigo para ler o seu resumo na totalidade e perder o interesse inicial.

Tipicamente, os SRs redirecionam uma pequena percentagem do tráfego (número de visitantes/utilizadores online) para diferentes motores de recomendação, e registam as interações dos utilizadores com os diferentes sistemas (Ricci et al. 2015).

As avaliações online são realizadas em ambientes controlados, o que permite identificar os efeitos de, por exemplo, alterações no algoritmo de recomendação nas métricas do utilizador. A metodologia normalmente adotada é os testes A/B, cujo objetivo é adquirir informação sobre a experiência de utilizador. Define-se, para tal, uma experiência aleatória com duas variantes, A e B. Os utilizadores participantes são segmentados a fim de determinar qual dessas variantes mais impacto significativo tem na experiência do utilizador com o SR e as suas recomendações.

Os registos de utilização são uma outra maneira de avaliar o SR online. Quanto se tem este tipo de registos, que indicam a forma como o utilizador realmente faz uso do sistema, pode-se descobrir de que forma as recomendações do SR se relacionam com os itens que ele selecionou. Isso é possível através da construção de um registo daquilo que foi mostrado ao utilizador (itens), no que ele clicou, o que ele comprou ou consumiu. Com essa informação pode-se modificar ou desenvolver algoritmos diferentes ao algoritmo de recomendação em avaliação.

O objetivo máximo da avaliação online é filtrar as abordagens inapropriadas. É, por isso, mais indicado realizar uma avaliação online por último, após um extenso estudo offline e/ou um estudo do utilizador. Reduz isto o risco de causar a insatisfação significativa do utilizador (Ricci et al. 2015). Uma boa prática seria planear avaliações online dos modelos de previsão do SR com melhor desempenho na avaliação offline.

Métricas de avaliação online

As taxas de aceitação são habitualmente medidas pelo rácio entre o número de recomendações clicadas pelo utilizador (para as aceder) e o número de recomendações apresentadas ao utilizador. Este rácio é denominado de taxas de cliques (*click-through rates* – CTR). Imagine-se que um SR exibe 8 000 recomendações. Se o utilizador clicar em 99 delas, então a CTR é de, aproximadamente, 1,2%. Uma qualidade positiva de métricas como a CTR é que elas podem ser uma medida explícita de eficácia.

Outra métrica de avaliação é a taxa de visita após recomendação (*visit after recommended rate*). Ela mede o número de vezes que o utilizador visitou um item recomendado após a primeira visita. Não obstante à crítica inicial sobre a insuficiência das métricas como único método de avaliação, elas complementam a avaliação online. Em especial, nas situações em que se pretende aferir se as discrepâncias entre o comportamento do utilizador e outras métricas de performance são estatisticamente significativas.

4.4.3 Estudos de utilizador

Os estudos de utilizador, tal como a avaliação online, é uma abordagem alternativa ou complementar às métricas tradicionais da avaliação offline. São uma forma de investigação que tanto pode ocorrer em laboratório ou no campo (Beel et al. 2016).

Nos estudos de laboratório, os participantes têm conhecimento sobre o seu papel na atividade, tendo consentido às condições do estudo (p.e. recolha de dados e partilha dos resultados junto da comunidade científica). O senão é que esta consciencialização pode afetar o comportamento dos utilizadores e, por consequência, comprometer os resultados da avaliação do SR em causa.

Nos estudos do mundo real, os participantes não estão cientes da sua participação no estudo. As classificações dadas pelos mesmos aos itens recomendados pelo SR não são, assim, influenciadas por qualquer tipo de informação. O ato tem por base o benefício próprio, uma vez que só com classificações honestas poderá o SR melhorar ou gerar recomendações, das quais a satisfação do utilizador depende.

Os estudos de utilizador medem a satisfação dos utilizadores de acordo com a qualidade apercebida dos mesmos, quer seja através de classificações explícitas (estudo quantitativo) ou *feedback* qualitativo (estudo qualitativo).

Em estudos quantitativos, os utilizadores recebem recomendações geradas por diferentes algoritmos de recomendação, classificam-nas, e a abordagem com a classificação média mais alta é considerada a mais eficaz. Os participantes são normalmente solicitados a quantificar a sua satisfação geral com as recomendações, contudo, também lhes pode ser pedido que classifiquem aspetos individuais do SR.

Nos casos em que é recolhido feedback qualitativo, o resultado do estudo de utilizador vincula-se às perguntas colocadas. Os participantes são convidados a navegar no SR em avaliação e as suas recomendações, selecionar um certo número de itens e classificá-los. Com base nas classificações, o sistema retorna outras recomendações, o participante explora os resultados e respondem a uma série de perguntas quanto à qualidade das recomendações (Kanakia et al. 2019). Através das respostas e comportamento durante a avaliação, podem ser calculadas as métricas presentes em 4.4.1.

4.4.4 Comparação dos algoritmos de recomendação e testes estatísticos

Após as avaliações, é pertinente realizar uma análise da significância estatística dos resultados observados. O intento é averiguar se a performance do algoritmo com melhor desempenho não se deve ao facto de o conjunto de dados escolhido se adequar de forma demasiado perfeita à sua abordagem.

Corre-se sempre o risco de os dados se encaixarem melhor num dos algoritmos do que nos outros quando estes não modelam de forma precisa a natureza dos utilizadores e dados reais.

Os testes de significância sustentam-se em medidas de evidência como o valor-p. O valor-p é calculado com base na suposição de que a hipótese nula é verdadeira para a população e que a diferença na amostra se deve ao acaso (Cañamares et al. 2020; Shani e Gunawardana 2011). Suponha-se que se quer determinar se:

H1: O algoritmo A é melhor que o algoritmo B para p=0,05.

A hipótese nula será o oposto de H1:

H0: O algoritmo A não é melhor que o algoritmo B para p=0,05

O p indica um limiar. Se o valor-p exceder p (valor-p > p), rejeita-se H0 e aceita-se H1. Se o valor-p for igual ou menor a p (valor p ≤ p), aceita-se H0 e rejeita-se H1.

Neste exemplo, se valor-p > 0,05, pode-se afirmar que se tem 95% (1,00 – 0,05 = 0,95 = 95%) de confiança que "H0: O algoritmo A não é melhor +que o algoritmo B para p=0,05" é falso. Por lógica, se H0 é falso, então "H1: O algoritmo A é melhor que o algoritmo B para p=0,05" deve ser verdadeiro.

Ao testar H0, quanto menor o resultado do valor- p, mais confiante se pode estar dos resultados obtidos na avaliação do SR.

O valor-p é uma medida utilizada em todas as estatísticas (testes-t, teste do sinal, teste de Wilcoxon, análises de regressão, etc). Sun et al. (2014, pág. 1340) testam a significância dos resultados de avaliação do método proposto através do valor-p de testes-t emparelhados.

Os testes-t analisam a diferença média entre as pontuações dos algoritmos A e B, normalizada pelo desvio padrão da diferença de pontuação. Em alternativa aos testes-t, tem-se valor-p do teste do sinal, conveniente para os casos em que se analisa qual a abordagem mais indicada para determinados utilizadores (Cañamares et al. 2020). Compara, portanto, o tamanho de dois grupos. Por exemplo, o tamanho ente o grupo 1, que é composto por utilizadores que preferem o algoritmo A, e o grupo 2, composto por utilizadores que preferem o algoritmo B. A hipótese nula do teste do sinal é "a diferença entre as medianas é zero".

4.4.5 Reprodutibilidade das experiências na avaliação de sistemas de recomendação

Os investigadores que realizam avaliações experimentais aos sistemas de recomendação devem descrever os seus procedimentos na totalidade, assegurando que outros possam replicar os seus resultados.

Nessa veia, é beneficial partilhar o conjunto de dados utilizado na avaliação offline de um SR, caso ele seja criado pelo autor da investigação ou modificado um *dataset* público pelo mesmo. Caso tal não se verifique, remove-se a possibilidade de o utilizar no futuro ou de replicar os experimentos que usufruíram deles. O mesmo se aplica aos conjuntos de dados disponibilizados por entidades, que estão sempre sujeitas ao término do seu projeto.

Algo notado ao longo da revisão da literatura foi a falta de menção do *framework* utilizado para a geração de recomendações e avaliação do sistema. O investigador pode criar um do zero ou servir-se de *frameworks* existentes como o LensKit[10], Mahout[11] (Barros et al. 2019; Beel et al. 2016), MyMediaLite[12], Duine[13], RecLab Core[14], easyrec[15], Recommender101[16] (Beel et al. 2016), e CF4J[17] (Barros et al. 2019).

Apesar de diferentes *frameworks* divergirem na implementação dos algoritmos, um entrave à comparação dos resultados da avaliação com *baselines*, é inusual serem feitas declarações sobre o *framework* utilizado nos artigos académicos.

Por questões de reprodutibilidade, é incentivado especificar:

- o *framework* utilizado e a versão;
- a configuração da avaliação (ex.: número de itens removidos do conjunto de dados e porquê,
- número de utilizadores, etc);
- a técnica utilizada para dividir o conjunto de dados;
- o tamanho do conjunto de dados para treino e para testes;
- o número de itens recomendados.

[10] https://lenskit.org/

[11] https://mahout.apache.org/

[12] http://www.mymedialite.net/

[13] http://www.duineframework.org/

[14] https://github.com/berkeley-reclab/RecLab

[15] https://github.com/hafael/easyrec-docker

[16] https://ls13-www.cs.tu-dortmund.de/homepage/recommender101/index.shtml

[17] https://github.com/ferortega/cf4j

5. PLANO DE TESTES AO SISTEMA DE RECOMENDAÇÃO CIENTÍFICA IVISSEM

O presente documento surge no âmbito do projeto de investigação *IViSSEM – 6.849,32 Journal Articles Everyday: Visualize or Perish!*, financiado pela Fundação para a Ciência e a Tecnologia. Ele introduz o plano de testes a aplicar ao sistema de recomendação IViSSEM no momento em que um protótipo do mesmo esteja concluído e pronto a ser avaliado.

A abordagem optada começa com uma avaliação offline, centrada no sistema. É seguida por um estudo de utilizador, uma abordagem centrada no utilizador e mais recurso-intensivo em comparação.

Os testes a realizar visam comparar os algoritmos de recomendação candidatos entre si ou comparar um único algoritmo de recomendação – proposto pelo IViSSEM – com *baselines* e verificar qual o melhor em duas medidas: (1) capacidade de prever a reação de um utilizador face as recomendações do sistema, (2) ordem das recomendações.

Apresenta-se a possibilidade de comparar algoritmos de recomendação candidatos ou de comparar a proposta de um só algoritmo de recomendação com *baselines*, mas apenas uma delas será seguida, conforme aquilo que tiver sido desenvolvido pelo IViSSEM.

Algoritmos de recomendação candidatos são algoritmos que definem o funcionamento de um sistema de recomendação, propostos por um autor e que, durante a avaliação, são comparados entre si para determinar e selecionar o melhor (de acordo com critérios previamente estabelecidos). São candidatos porque apenas um deles será escolhido como o mais adequado. *Baselines* são algoritmos de referência propostos na literatura. As *baselines* podem ser abordagens representativas do estado da arte e/ou abordagens propostas por terceiros.

5.1 Descrição do projeto

"Mais de 2,5 mil milhões de artigos científicos são publicados anualmente. Mais de 6.849,32 novos artigos de revistas de investigação são publicados todos os dias! Como é que no mundo um investigador encontra o que importa?

A mera identificação dos *Scientific Knowledge Objects* (SKOs) mais relevantes num determinado tópico é cada vez mais difícil devido às interfaces existentes, devolvendo listas massivas de resultados. É

reconhecido que os investigadores não são meros produtores de conhecimento. Em vez disso, são atores sociais que desempenham um papel preponderante na descoberta e filtragem do conhecimento científico. Os dados resultantes desta interação social fornecem uma base importante para a conceção de várias métricas de utilização, também conhecidas como *altmetrics*.

O acesso à informação correta e relevante é primordial para as descobertas científicas. O IViSSEM visa desenvolver e testar uma nova *altmetric*, chamada Social *Scholarly Experience Metric*. Esta métrica resultará da aplicação de técnicas de *Machine Learning* a diferentes combinações de *altmetrics* e perfis de investigadores. A sua aplicação irá refletir as preferências individuais no processo de encontrar um tópico específico. As atuais listas maciças de resultados serão substituídas por uma interface inovadora baseada em técnicas avançadas de visualização" (IViSSEM, sem data).

5.2 Objetivos

Os testes a implementar assumem duas vertentes:

1. Comparar os algoritmos de recomendação candidatos, ou comparar um algoritmo de recomendação com *baselines*, e verificar qual o melhor nas seguintes medidas:
    a. capacidade de prever a reação de um utilizador face as recomendações do sistema de recomendação – a interface do SR IViSSEM permite ao utilizador aceitar/rejeitar recomendações e classificar o seu nível de satisfação com as recomendações adicionadas à biblioteca numa escala definida por 3 emojis.
    b. ordem das recomendações - idealmente, a qualidade das recomendações segue uma ordem decrescente, ou seja, as recomendações mais relevantes encontram-se no topo da lista. No caso do SR do IViSSEM, a interface exibe essa ordem num gráfico *treemap*, em que o tamanho dos retângulos que o compõem corresponde à relevância da recomendação

    Perante os resultados, e caso se compare algoritmos de recomendação candidatos, seleciona-se aquele com melhor desempenho nas medidas acima mencionadas. Caso se trate de uma comparação entre a proposta de um único algoritmo de recomendação e *baselines*, os resultados demonstram se o algoritmo proposto é superior ou equiparável a algoritmos de recomendação do estado da arte, e em que medidas.

2. Testar a interface do sistema de recomendação e averiguar se:

a. Os utilizadores conseguem completar um conjunto de tarefas específicas com sucesso – que determinará se a interface contém a informação necessária para o utilizador navegar na plataforma com facilidade e que ajustes ou mudanças terão que ser feitas para melhorar a experiência do utilizador.
b. os utilizadores estão satisfeitos com a interface a nível estético.

5.3 Estratégia e abordagem dos testes

Para atingir os objetivos, optar-se-á por uma avaliação offline, seguida de um estudo de utilizador. Uma avaliação offline é centrada no sistema e um estudo de utilizador é centrado no utilizador.

A avaliação offline é normalmente mais fácil de realizar por não requer interação com utilizadores reais nem exigir a implementação do SR, o que significa que não é recursos-intensivo.

Os estudos de utilizador envolvem, geralmente, um pequeno número de participantes, incumbido de realizar tarefas específicas num SR. O mais benéfico desta abordagem é que permite obter feedback em áreas como a performance do algoritmo e interface do sistema.

Enquanto que a primeira abordagem simula a utilização do SR, as abordagens centradas no utilizador permitem observar o SR no mundo real, sendo por isso preferível. No entanto, as abordagens complementam-se. Recomenda-se iniciar pela avaliação offline e, de seguida, realizar o estudo de utilizador devido aos recursos que este último requer.

No momento de concretização da avaliação offline, primeiro colocam-se os algoritmos candidatos a funcionar, ou o algoritmo de recomendação proposto e as *baselines*. De seguida, geram-se recomendações para os utilizadores do sistema e, por fim, aplicam-se métricas de avaliação de desempenho. Consoante os resultados obtidos com as métricas, seleciona-se o algoritmo de recomendação com melhor desempenho ou compara-se o desempenho do algoritmo proposto com o desempenho das *baselines*.

O algoritmo candidato selecionado na etapa final do processo da avaliação offline, ou o único algoritmo de recomendação proposto, é aplicado a uma interface funcional e testável. Essa interface corresponde, neste caso, ao website do sistema de recomendação do IViSSEM onde os utilizadores podem navegar por entre as recomendações geradas e dar feedback em relação a elas. Com base nas caraterísticas da interface, indivíduos distanciados do projeto são convidados a participar numa atividade onde eles têm de realizar determinadas tarefas e preencher um questionário sobre a experiência. O questionário é posteriormente analisado.

No contexto de se efetuar a avaliação a um só algoritmo e este ter uma performance inferior às *baselines*, tem-se duas vias: 1) do algoritmo derivam-se algoritmos candidatos com o objetivo de se construírem outros mais eficientes e eficazes ou 2) procede-se com o estudo de utilizador com o objetivo de verificar se existem disparidades entre as ambas avaliações, uma vez que o estudo de utilizador reflete o mundo real, ao contrário da avaliação offline.

5.4 Critérios para a realização dos testes

Para realizar a avaliação offline é necessário:

- que o código do algoritmo de recomendação proposto e as *baselines*, ou os algoritmos de recomendação candidatos, esteja pronto a correr e, portanto, avaliar.
- um *dataset* compatível com os dados com que o sistema de recomendação irá lidar (na vida real e no estudo de utilizador).

No caso do estudo de utilizador, é necessário:

- um protótipo do SR IViSSEM, pronto a ser utilizado por indivíduos reais que vão executar tarefas específicas durante o estudo de utilizador.

5.5 Metodologia da avaliação offline

Esta secção apresenta o procedimento para avaliar a eficácia do ou dos algoritmos de recomendação.

5.5.1 Requisitos para a realização da avaliação offline

1. *Python e Integrated Development Environment* (IDE) – onde o código está escrito, é testado e depurado (*debugged*).
2. Biblioteca Scikit-learn[18] e scikit Surprise[19] – Scikit-learn é uma biblioteca de *machine learning* para Python. Permite construir experiências reprodutíveis. Oferece implementações de *collaborative filtering*, rotinas de preparação de dados, ferramentas para executar algoritmos de

[18] https://scikit-learn.org/

[19] http://surpriselib.com/

recomendação em série e métricas de avaliação. As caraterísticas da biblioteca podem ser consultadas na documentação oficial[20]. Surprise é um scikit para construir e avaliar SRs que lidam com dados de classificação explícitos

3. *Dataset* (com items, utilizadores, classificações) – devido ao inacesso a dados do mundo real, o sistema de recomendação é avaliado offline através de um *dataset*.

5.5.2 Antes da avaliação offline

Algoritmos de recomendação candidatos e baselines

O presente documento reconhece a possibilidade de se aplicar a avaliação offline a um conjunto de algoritmos candidatos ou a um único algoritmo de recomendação. No último caso, é necessário que se tenham as *baselines* definidas. Esta secção apresenta *baselines* a utilizar na eventualidade de este ser o percurso optado para avaliar o SR.

- *Singular Value Decomposition* (SVD) (Zhang et al., 2005): utiliza o método SVD com base na matriz de classificação do utilizador para encontrar um modelo de baixa dimensão que maximize a probabilidade das classificações observadas nos sistemas de recomendação. É uma técnica de recomendação baseada em *collaborative filtering*. Encontra-se presente em Surprise.
- *Scholarly Paper Recommendation via User's Recent Research Interests* de Sugiyama et al. (2010): propõem recomendar trabalhos académicos relevantes através da captura dos interesses do investigador através das suas publicações passadas. A representação do perfil do utilizador incorpora publicações passadas e documentos vizinhos (citações e referências).
- A hybrid approach for article recommendation in research social networks de Sun et al. (2017): na primeira fase, estabelece-se uma correspondência entre o perfil do utilizador e o perfil de artigos científicos. Na segunda fase, a relevância, uma pontuação de conectividade e uma pontuação de qualidade são agregadas à distribuição de ponderação, gerando um ranking de recomendações. É uma técnica de recomendação híbrida de *content-based filtering* e *collaborative filtering*.

[20] https://surprise.readthedocs.io/en/stable/

Dataset

Numa avaliação offline é possível selecionar um *dataset* disponível online ou, caso não se encontre um *dataset* compatível com os dados com que o sistema lidará, poder-se-á construir um *dataset* próprio (total ou parcialmente) com o auxílio de um *web crawler* e/ou a um gerador de dados.

Para o caso do presente plano de testes, recomenda-se o *dataset* de Sugiyama e Kan (2015)[21]. O *dataset* contém uma lista de publicações de 50 investigadores cujos interesses de investigação são de diversos campos da informática (recuperação de informação, engenharia de software, interface de utilizador, segurança, gráficos, bases de dados, sistemas operativos, sistemas incorporados e linguagens de programação). A figura 13 mostra algumas estatísticas do *dataset*.

Na eventual necessidade de produzir dados de classificação entre utilizadores e ROs, tal pode ser feito com base nas relações artigo-citação, em que se assume que uma citação corresponde a uma recomendação.

Figura 13 Estatísticas do dataset de Sugiyama e Kan (2015).

Total number of researchers	50
Average number of researchers' publications	10
Average number of citations of each researchers' publications	14.8 (max. 169)
Average number of references to each researchers' publications	15.0 (max. 58)
Total number of recommending papers	100,351
Average number of citations of the recommending papers	17.9 (max. 175)
Average number of references to the recommending papers	15.5 (max. 53)

https://doi.org/10.1371/journal.pone.0184516.t001

Fonte: Haruna et al. (2017)

Na salvaguarda de ser necessário utilizar um outro *dataset* por este se revelar impossível ou complicado de utilizar, menciona-se ainda o ArnetMiner[22] de Tang et al. (2008), com informação sobre itens (publicações/artigos), citações, autores e colaborações de autor; os datasets de Sugiyama e Kan (2013).

[21] https://scholarbank.nus.edu.sg/handle/10635/146027

[22] https://www.aminer.org/aminernetwork

Instalação de Scikit-learn e Surprise

Esta biblioteca irá suportar a realização das tarefas da avaliação offline. Para instalar a biblioteca Scikit-learn indica-se o site oficial, onde as instruções podem ser lidas e consultadas[23]. Também o website de Surprise tem devidas instruções de instalação[24].

Antes de se proceder com a implementação do que é prescrito nos próximos capítulos, deve-se ler com atenção a página web de Surprise Getting Started[25], que fornece uma visão simples de como aplicar a biblioteca na realização da avaliação offline.

Esta biblioteca irá auxiliar, em especial, o processo de divisão do *dataset* em dados para treino do modelo de recomendação e em dados para teste, na comparação entre os algoritmos, e na aplicação de métricas para avaliar o desempenho dos algoritmos de recomendação candidatos.

5.5.3 Durante a avaliação offline

Carregamento dos algoritmos de recomendação em Surprise

O programa Surprise já inclui certos algoritmos de recomendação, contudo o algoritmo do IViSSEM é original, pelo que se deve seguir o modelo de algoritmos de Surprise, que indica como criar/carregar um algoritmo de recomendação próprio. Esta ação pode ser feita consultando o capítulo H*ow to build your own prediction algorithm*[26]da documentação oficial.

Carregamento e divisão do dataset em dados para treino e para teste

Para carregar o *dataset* em Surprise, deve-se seguir as instruções da página web *Getting Started*[27] e a página *dataset* module[28] da documentação oficial.

O *dataset* carregado é dividido em conjunto de dados para treino e conjunto de dados para teste. O conjunto de dados para treino é utilizado para gerar o modelo de recomendação e o conjunto de dados para teste é utilizado para avaliar o desempenho do modelo.

---

[23] http://scikit-learn.org/stable/install.html

[24] http://surpriselib.com/

[25] https://surprise.readthedocs.io/en/stable/getting_started.html

[26] https://surprise.readthedocs.io/en/stable/building_custom_algo.html

[27] https://surprise.readthedocs.io/en/stable/getting_started.html#use-a-custom-dataset

[28] https://surprise.readthedocs.io/en/stable/dataset.html

Existem várias estratégias para dividir o *dataset*, de entre as quais, a validação cruzada *k-fold*. k é um parâmetro que representa o número de subconjuntos em que um conjunto de dados se divide. Uma vez feita essa divisão, uma parte desse subconjunto pode ser utilizado como conjunto de dados para teste e a parte restante é utilizada como conjunto de dados para treino.

O procedimento para a validação cruzada *k-fold* é o seguinte:

1. Baralhar o *dataset* aleatoriamente
2. Dividir o *dataset* em k subconjuntos
3. Para cada subconjunto:
    a. Selecionar uma parte como conjunto de dados para teste
    b. Selecionar a parte restante como conjunto de dados para treino
    c. Treinar o modelo de recomendação nos conjuntos de treino e avaliar o mesmo modelo nos conjuntos de teste
4. Avaliar o desempenho do modelo de recomendação com base na média dos resultados das k avaliações.

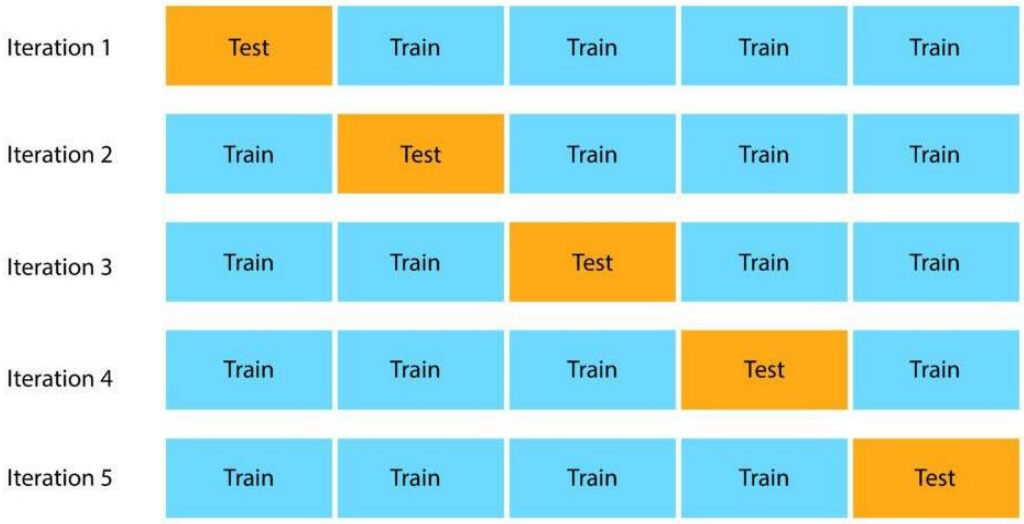
Para o sistema de recomendação do projeto IViSSEM prescreve-se uma validação cruzada *5-fold* (figura 14), que consiste em dividir o (número de observações no) *dataset* em 5 subconjuntos, depois de este ter sido baralhado de modo aleatório. Cada um dos subconjuntos deve ter o mesmo número de observações.

Para os 5 subconjuntos, 20% (das observações) constituirá o conjunto de dados de teste e 80% constituirá o conjunto de dados de treino. Por outras palavras, 1 dos 5 subconjuntos representa o conjunto de dados de teste e os restantes 4 dos 5 subconjuntos representam o conjunto de dados de treino. No total, o modelo é testado em cinco iterações:

- Na iteração 1, o modelo é treinado com o subconjunto 1, subconjunto 2, subconjunto 3 e subconjunto 4, e testado com o subconjunto 5.
- Na iteração 2, o modelo é treinado com o subconjunto 2, subconjunto 3, subconjunto 4 e subconjunto 5, e testado com o subconjunto 1.
- Na iteração 3, o modelo é treinado com o subconjunto 1, subconjunto 3, subconjunto 4 e subconjunto 5, e testado com o subconjunto 2.
- Na iteração 4, o modelo é treinado com o subconjunto 1, subconjunto 2, subconjunto 4 e subconjunto 5, e testado com o subconjunto 3.

- Na iteração 5, o modelo é treinado com o subconjunto 1, subconjunto 2, subconjunto 3 e subconjunto 5, e testado com o subconjunto 4.

Figura 14 Representação de uma validação cruzada 5-fold.



Fonte: Great Learning Team (2020, setembro 24).

A validação cruzada *5-fold* prescrita é feita com a Biblioteca Scikit-Learn e Surprise, seguindo a página web de Surprise *Getting Started* e o subcapítulo *Use cross-validation iterators*[29]. Em adição, recomenda-se igualmente o capítulo *The model_selection package*[30] da documentação oficial.

Métricas a aplicar na avaliação

A etapa seguinte é a avaliação do desempenho do sistema de recomendação. Ele é aferido através do uso de métricas de avaliação, nomeadamente métricas de precisão preditiva e métricas de precisão top-N. Elas são aplicadas aos resultados do conjunto de dados de teste.

Para cada algoritmo candidato ou, algoritmo e *baselines*, são aplicadas as métricas para que se possam fazer comparações e elações.

A aplicação das métricas é realizada com Surprise, seguindo as indicações da documentação oficial, em especial o capítulo *accuracy module*[31]e o subcapítulo *How to compute precision@k and recall@k*[32].

[29] https://surprise.readthedocs.io/en/stable/getting_started.html#use-cross-validation-iterators

[30] https://surprise.readthedocs.io/en/stable/model_selection.html

[31] https://surprise.readthedocs.io/en/stable/accuracy.html

[32] https://surprise.readthedocs.io/en/stable/FAQ.html#how-to-compute-precision-k-and-recall-k

As métricas devem ser calculadas @k (precisão@k, recall@k, MAP@k). Por exemplo, a precisão@k é a ração de itens relevantes nas recomendações do top k. Num momento inicial decidiu-se que o SR IViSSEM geraria um top k de 20, mas durante a avaliação sugere-se fazer k=5, 10, 15 e 20 para verificar se, de facto, recomendar os top 20 ROs é o mais indicado. Surprise ajuda neste sentido.

As métricas que se apresenta na tabela 14 são as que devem ser aplicadas ao algoritmo ou algoritmos candidatos do sistema de recomendação IViSSEM.

Tabela 14 Métricas a aplicar ao sistema de recomendação IViSSEM.

Métricas de tomada de decisão		
Equação	Descrição	Finalidade
$Precisão = \frac{Nº\ de\ ROs\ relevantes\ recomendados}{Nº\ total\ de\ ROs\ recomendados}$	Fração de ROs relevantes, recomendados a um investigador.	Medir a capacidade do algoritmo em recomendar ROs relevantes nas primeiras k recomendações.
$Recall = \frac{Nº\ de\ ROs\ relevantes\ recomendados}{Nº\ total\ de\ ROs\ relevantes}$	Fração de ROs relevantes para o número total de ROs relevantes no sistema.	Medir a capacidade do algoritmo em identificar ROs relevantes.
$F1 = \frac{2 \times Precisão\ \times Recall}{Precisão\ + Recall}$	Média harmónica da precisão e do *recall*.	Medir o número de casos classificados incorretamente pelo algoritmo.
Métricas de precisão top-N		
Equação	*Descrição*	*Finalidade*
$MAP = \frac{1}{U}\sum_{i=1}^{U}\frac{1}{m_j}\sum_{k=1}^{N}P(R_{jk})$ K é o número de ROs recomendados; N é o tamanho da lista de recomendação; U é o número total de utilizadores registados no SR; m_j é o número de ROs relevantes ao utilizador j na lista de recomendação; $P(R_{jk})$ é a precisão dos resultados obtidos a partir do primeiro resultado até chegar ao RO k.	Avalia a qualidade do ranking das recomendações.	Avaliar as recomendações aceites/rejeitadas pelo utilizador.
$nDCG_k = \frac{DC_k}{IDCG_k}$ K é o número de ROs recomendados; DCG é o *discounted cumulative gain* das recomendações; IDCG é o *Ideal Discounted Cumulative Gain.*.	Avalia a qualidade do ranking das recomendações.	Avaliar as recomendações aceites, adicionadas à biblioteca pessoal e classificadas pelos utilizadores.

5.5.4 Depois da avaliação offline

Comparação dos resultados das métricas

Obtidos os resultados das métricas (apresentadas no capítulo anterior) de cada algoritmo candidato ou do único algoritmo de recomendação e *baselines*, deve-se calcular o valor-p com o Excel ou semelhante, para estudar a significância destes. Se, por exemplo, valor-p > 0,05, pode-se afirmar que se tem 95% (1,00 – 0,05 = 0,95 = 95%) de confiança que "H0= O algoritmo A não é melhor que o algoritmo B para p=0,05".

De seguida, é necessário comparar os resultados das métricas para determinar qual o algoritmo com melhor desempenho. A comparação deve ser efetuada com o auxílio de Surprise[33].Para cada métrica, essa comparação deve ser feita de acordo com os seguintes critérios:

- Os valores da precisão, *recall*, e F1 variam entre 0 e 1. Tem melhor desempenho o algoritmo com os valores mais próximos de 1.
- O valor do MAP varia entre 0 e 1. Tem melhor desempenho o algoritmo com um valor mais próximo de 1.

Comparados e analisados os valores obtidos nas métricas, seleciona-se o algoritmo de recomendação candidato com melhor desempenho no geral. Ou, no caso da avaliação de um algoritmo de recomendação com *baselines*, idealmente, o algoritmo proposto tem um melhor desempenho no geral que as *baselines*.

Outro output que deve advir da avaliação offline é o @k que o sistema de recomendação deve incorporar. Isto é, com os resultados das métricas, pode-se observar qual o número de recomendações que o sistema deve apresentar ao utilizador para um melhor desempenho.

Procede-se o plano de testes com o estudo de utilizador. Num estudo de utilizador e possível optar-se por avaliar o algoritmo de recomendação com melhor desempenho na avaliação offline, ou o único algoritmo proposto, ou todos os algoritmos candidatos.

No entanto, na eventualidade de se efetuar a avaliação a um só algoritmo e este ter uma performance inferior às *baselines*, tem-se duas vias: 1) do algoritmo derivam-se algoritmos candidatos com o objetivo de se construírem outros mais eficientes e eficazes ou 2) procede-se com o estudo de utilizador com o

[33] https://nbviewer.org/github/NicolasHug/Surprise/blob/master/examples/notebooks/Compare.ipynb

objetivo de verificar se existem disparidades entre as ambas avaliações, uma vez que o estudo de utilizador reflete o mundo real, ao contrário da avaliação offline. O caminho a tomar depende de uma reflexão da circunstância.

5.6 Metodologia do estudo de utilizador

Esta secção apresenta o guião para avaliar um só algoritmo de recomendação, mas na ótica do utilizador. O primeiro objetivo do estudo de utilizador é calcular o desempenho do algoritmo na vida real.

Como mencionado brevemente no capítulo anterior, o estudo de utilizador pode ser implementado para um único algoritmo candidato, dois ou mesmo todos os algoritmos candidatos. Para o presente plano de testes optou-se por descrever o estudo para somente um algoritmo, mas ele pode ser facilmente adaptado para vários, uma vez que os passos a seguir são os mesmos.

Outra nota importante de referir é que, ao estudo de utilizador, se recomenda acrescentar um segundo objetivo: testar a interface do sistema de recomendação. Assim, o plano para este estudo é completado pelo plano de testes para a usabilidade do design, delineado por Bruno Azevedo, quem integrou o projeto IViSSEM. Nesse sentido, a estratégia que de seguida se apresenta é uma conjugação do primeiro e segundo objetivo, conciliada com o documento de *design usability* de Bruno Azevedo. Dessa forma, a partir desta fase deverá ser levado em conta o documento mencionado.

O estudo de utilizador é executado recrutando um conjunto de participantes voluntários, aos quais se pede que realizem várias tarefas, que exigem uma interação com o sistema de recomendação e interface. Ele é realizado em 3 rondas/sessões, à semelhança do teste de *design usability*. No que diz respeito ao primeiro objetivo proposto, as 2 primeiras rondas servem para treinar o algoritmo e ajustar as recomendações às preferências de cada participante, e a última ronda serve para testar efetivamente o algoritmo.

Nesse sentido, serão convidados indivíduos não familiarizados com a plataforma de recomendação. Depois de obtido o consentimento expresso de cada um, os participantes serão convidados a navegar o sistema de recomendação, a explorar e classificar as recomendações geradas e a partilhar o seu parecer acerca da interface e das recomendações através de um questionário.

Em adição ao questionário, os dados gerados durante a utilização do sistema (e.g. páginas visitadas ou links clicados) devem ser registados e o ecrã do utilizador deve ser gravado durante a realização do teste. Isto permitirá observar posteriormente o comportamento do participante e detetar possíveis problemas

de experiência de navegação. Também a partir dos dados de navegação dos participantes durante a avaliação, podem ser calculadas as métricas precisão, *recall*, F1 e MAP.

O estudo de utilizador deve ser realizado em rondas e cada ronda pode ter de ser realizada em horários diferentes para grupos de participantes conforme a capacidade do espaço onde o estudo se irá realizar. De relembrar que para o estudo de utilizador ser possível, é necessário que haja um protótipo da interface do sistema de recomendação IViSSEM pronto a ser utilizado pelos participantes do estudo – a interface não tem necessariamente de estar num estado final, mas num estado que permita ao participante realizar as tarefas que serão especificadas nas páginas que se seguem.

5.6.1 Requisitos para a realização do estudo de utilizador

Previamente à realização do estudo de utilizador os aspetos que se seguem devem estar desenvolvidos e preparados a serem utilizados:

- Participantes voluntários – metade atua como iniciantes num determinado domínio de conhecimento e a outra metade atua como peritos num determinado domínio de conhecimento,
- Espaço físico – para realizar o estudo, por exemplo, a sala dedicada ao projeto IViSSEM no Campus de Couros,
- Formulário de consentimento informado (apêndice II) assinado pelos participantes,
- Software de gravação de ecrã,
- Base de dados associada à interface do sistema de recomendação que armazene a interação dos utilizadores com a mesma,
- Tarefas a realizar pelo utilizador em interação com a interface definidas,
- Folha com indicação das tarefas a realizar para distribuir pelos participantes,
- Questionário pós-experiência a preencher pelos participantes (apêndice V).

5.6.2 Antes do estudo de utilizador

Número de participantes

Para avaliar a interface, o plano de testes de usabilidade de design descreve serem necessários 3 a 5 participantes para cada 1 de 3 rondas. Contudo, para a avaliação do sistema de recomendação, o mesmo não é adequado. Na revisão de literatura de Beel et al. (2016) sobre a avaliação de sistemas de recomendação académicos, os autores denotam que 60% das abordagens avaliadas com um estudo de utilizador, 60% tinham 15 ou menos participantes ou não revelaram o número de participantes. No

entanto, concluem que mais de metade dos estudos de utilizador não recrutam um número significativo de participantes. Dos restantes estudos, 17% deles utilizam 16 a 50 participantes e 25% utilizam mais de 50 participantes. Apesar destas observações, os autores da revisão de literatura não prescrevem uma quantidade mínima de participantes.

Irá ser necessário conciliar o número de participantes para o teste de usabilidade e para o teste do sistema de recomendação, ainda que ambos possam ser feitos no mesmo dia e os participantes do teste de usabilidade possam ser igualmente participantes no teste do sistema de recomendação. A diferença é que terão tarefas acrescidas.

Seleção dos participantes

Os participantes selecionados devem ser i. externos ao projeto, e ii. Investigadores, académicos, estudantes ou outro, desde que a ocupação profissional esteja intimamente conectada à procura e uso de research outputs. Em adição, iii. metade dos participantes deve possuir elevado conhecimento num domínio científico de ciências de computação ou de sistemas de informação ou similares (ou seja, peritos num domínio compatível com o domínio dos *research outputs* presentes na base de dados do sistema de recomendação), e iv. a outra metade dos participantes deve possuir baixo conhecimento num domínio científico de ciências de computação ou de sistemas de informação ou similares. Com esta decisão, pretende-se averiguar se existe alguma discrepância na qualidade das recomendações, consoante o nível de conhecimento do utilizador do sistema.

No processo de seleção dos participantes, deve-se recolher informação sobre:

1. Idade,
2. Género,
3. Ocupação profissional,
4. Nível de conhecimentos no domínio científico de ciências de computação, sistemas de informação e afins,
5. Frequência de utilização de sistemas de recomendação científica ou similares.

Esta informação serve não só para o propósito da seleção, mas também para fins estatísticos.

O nível de conhecimento define os dois principais grupos de participantes, no entanto, na altura de analisar os resultados do estudo de utilizador pode-se averiguar se existe alguma relação entre os grupos de participantes ao olhar para frequência com que eles utilizam sistemas de recomendação ou similares.

Teste-piloto das atividades

Um teste-piloto deve ser realizado antes das sessões do estudo de utilizador. O objetivo é detetar falhas antecipadamente e, se necessário, redesenhar elementos da interface ou dos requisitos. Garante isto que o estudo está em condições de ser implementado.

O participante do teste-piloto pode ser um membro do projeto, mas desconectado do desenvolvimento do teste ou do (interface do) sistema. Em alternativa, o participante pode ser um colega com pouco ou nenhum conhecimento do projeto.

Consentimento informado

Aos potenciais participantes devem ser enviadas informações sobre a natureza do projeto (objetivo do projeto e objetivo do estudo de utilizador) e detalhes relacionados com a sua participação na atividade em questão para que possam tomar uma decisão informada. Contactos devem ser incluídos para potenciais dúvidas.

Um formulário deverá ser assinado pelos participantes, explicitando que o participante:

1. consente a participar no estudo
2. foi informado com antecedência
3. foi informado dos procedimentos do estudo
4. foi informado das gravações e do tratamento das mesmas
5. consente à coleta dos dados gravação do ecrã e utilização do sistema
6. está informado aos contactos a utilizar em caso de dúvidas

Caso o estudo de utilizador tenha de ser autorizado por uma entidade especifica, é possível que o documento de consentimento informado seja fornecido pela mesmo. Na eventualidade de tal não acontecer, indica-se o apêndice II, que consiste num formulário modelo passível de ser modificado.

5.6.3 Durante o estudo de utilizador

Nas sessões do estudo de utilizador, os participantes devem ser relembrados do propósito do projeto e, principalmente, do estudo de utilizador. Depois de contextualizados, é importante explicar-lhes o que irá ser pedido deles.

Deve-se pedir aos participantes, antes do início da atividade, que utilizem a interface e interajam com o sistema de recomendação, seguindo as tarefas da tabela 15. Aos participantes deve ser distribuída uma folha com o número e descrição das tarefas a executar (coluna 1 e 2 da tabela 15, apêndice III).

A tabela, para além do número e descrição das tarefas, possui informação sobre a página da interface do sistema onde a tarefa deve ser realizada, o objetivo e a prioridade de cada tarefa. Estas colunas servem para analisar a performance dos participantes e de que maneira ela se relaciona com problemas de usabilidade ou outro.

A tabela é idêntica àquela que se encontra no plano de usabilidade, porém as tarefas 1.1 a 5.3 pertencem ao plano de teste de usabilidade enquanto que as tarefas 6.1 e 6.1.2 são exclusivas à validação do algoritmo de recomendação.

Para o caso da avaliação do algoritmo de recomendação, as últimas duas tarefas são suficientes. As tarefas anteriores servem, no entanto, para o participante compreender os mecanismos da interface para poder utilizá-la com reduzida dificuldade durante a atividade. Caso se opte por um *overlap* entre os participantes no teste de usabilidade e no teste do algoritmo, então a tabela 15 deve ser utilizada para esses participantes. Caso não, a tabela 15 pode ser também usada ou, em alternativa, realizam-se apenas as tarefas 6.1 e 6.1.2 depois de um rápido tutorial de como utilizar a interface.

Concluídas as tarefas, os participantes devem preencher um questionário, adaptado do trabalho de Pu et al. (2011). As perguntas recolhem dados que não são diretamente observáveis. Os autores desenvolveram o modelo ResQue *(Recommender systems' Quality of user experience),* que tem como objetivo avaliar as qualidades percebidas dos recomendadores, por exemplo, a usabilidade, utilidade, qualidades de interface e interação, satisfação dos utilizadores, e a influência destas qualidades nas intenções comportamentais dos utilizadores.

O questionário deve ser respondido na Escala de Likert de 1 a 5, em que:

- 1 corresponde a "Discordo totalmente",
- 2 corresponde a "Discordo",
- 3 corresponde a "Nem concordo nem discordo",
- 4 corresponde a "Concordo",
- 5 corresponde "Concordo totalmente".

Tabela 15 Tarefas a serem realizadas pelos participantes do estudo de utilizador.

Nºda tarefa	Descrição da tarefa	Página da interface onde a tarefa deve serrealizada	Objetivo	Prioridade
1.1	Crie uma conta.	*Landing page*	O utilizador deve ser capaz decriar uma conta sem dificuldades.	Baixa
1.2	Preencha o seu perfil com osseus domínicos de conhecimento.	*My Account*	O utilizador deve ser capaz de adicionar domínios de conhecimento ao seu perfil semqualquer ajuda.	Alta
2.1	Consulte os detalhes(metadados) de um artigocientífico.	*All Pages*	O utilizador deve poder visualizar os metadados de um artigo científico através da.*tooltip.*	Média
2.2.1	Rejeite uma recomendação ao interagir com o gráfico na página inicial.	*Overview*	O utilizador deve poder visualizaros metadados de um artigo científico através da *tooltip.*	Alta
2.2.2*	Aceite uma recomendação.	*Overview*	O utilizador deve ser capaz de encontrar facilmente o botão de aceitação.	Baixa
2.3	Adicione uma recomendaçãoà sua biblioteca pessoal.	*Overview*	O utilizador deve poder acrescentar uma recomendaçãoà "*My library*".	Média
2.3.1	Abra um artigo científico comose o fosse ler.	*My Library*	O utilizador deve saber que tem de clicar no link DOI para ler um artigo científico.	Alta
2.3.2	Recomende o artigo científicoque abriu.	*My Library*	O utilizador deve poder encontrara classificação/ recomendação emojis na *My Library* para recomendar o artigo científico que acrescentou à sua Biblioteca.	Alta
2.4.2.2	Rejeite uma recomendação através do menu/painel da página *Overview.*	*Overview/Metadata Panel*	O utilizador deve ser capaz de encontrar o botão de rejeição através do painel de metadados.	Alta
3.1	Na visualização *time travel*/filtre as recomendações pordata.	*Overview / time travel*	O utilizador deve ser capaz de encontrar o botão de filtrar as recomendações por data.	Média
4.1	Adicione um artigo científico à plataforma.	*Upload icon, Add on*	O utilizador deve ser capaz de adicionar um artigo científico à plataforma.	Alta
5.1	Procure por um recomendador e explore operfil deste.	*Recommenders*	O utilizador deve ser capaz de encontrar o perfil de um recomendador através da página*Recommenders.*	Baixa
5.2	Dê *follow* a um recomendador	*Recommenders*	O utilizador deve ser capaz deseguir um recomendador.	Baixa
5.3	Adicione uma recomendaçãode um recomendador – que esteja a seguir ou outro.	*Recommenders*	O utilizador deve ser capaz de interagir com as recomendaçõesdos recomendadores.	Média
6.1	Na página inicial, para cadaumas das recomendações, aceite ou rejeite a recomendação conforme o seu interesse pessoal.	*Overview*	Esta tarefa serve para avaliar o desempenho do algoritmo do sistema de recomendação na vida real.	Alta
6.1.2	Na página *My Library* aparece-lhe as recomendações aceites por si. Tendo lido, pelo menos, o resumo de cada artigo científico, classifique-os conforme o seu parecerpessoal, clicando nos emojis.	*My Library*	Esta tarefa serve para avaliar o desempenho do algoritmo do sistema de recomendação na vida real.	Alta

Fonte: Bruno Azevedo e a autora.

O questionário corresponde ao apêndice IV. A folha que deve ser distribuída aos participantes corresponde ao apêndice V. As respostas devem ser agregadas e guardadas numa folha de cálculo com um formato semelhante à tabela 16.

Tabela 16 Tabela exemplificativa das respostas agregadas ao questionário ResQue.

Participante	**Pergunta 1**	**Pergunta 2**	**Pergunta 3**	**...**	**Pergunta n**
1					
2					
3					
...					
n					

5.6.4 Depois do estudo de utilizador

Os dados recolhidos durante o estudo são analisados:

1. Com o armazenamento dos dados de navegação no protótipo, é possível calcular as métricas: precisão, *recall*, F1 e MAP, nDCG.
    a. Comparar com os resultados da avaliação e calcular a correlação entre ambos.
2. Através do questionário fazer aferências quanto à significância estatística das respostas e correlacioná-las:
    a. Calcular a medida de tendência central moda. Visualizar a distribuição das respostas num gráfico.
    b. Proceder com a técnica de inferência Mann Whitney para comparar as respostas dos participantes iniciantes num domínio de conhecimento com as respostas dos participantes peritos num domínio de conhecimento. A técnica pode ser efetuada no Excel, seguindo as indicações de Zaiontz (sem data).
        i. Para realizar o teste de Mann Whitney 3 pressupostos têm de ser cumpridos:
            1. A variável a analisar é ordinal ou contínua (e.g. escala Likert)
            2. As observações dos dois grupos são independentes um do outro
            3. A forma das distribuições para os dois grupos é idêntica
        ii. Procedimento:
            1. Testar as hipóteses
                a. H0: as duas populações são iguais

b. H1: as duas populações não são iguais

2. Determinar o nível de significância a usar para as hipóteses, por exemplo, 0.01, 0.05 e 0.1.
3. Encontrar a estatística do teste U

$$U_1 = n_1 \times n_2 + \frac{n_{1(n_1+1)}}{2} - R_1 \quad (22)$$

$$U_2 = n_1 \times n_2 + \frac{n_{2(n_2+1)}}{2} - R_1 \quad (23)$$

Onde $n1$ e $n2$ são tamanhos de amostra para as amostras 1 e 2, e R_1 e R_2 são a soma das classificações para as amostras 1 e 2, respetivamente.

4. Rejeitar ou aceitar a hipótese nula com base em 3.
5. Interpretar os resultados

c. Para encontrar as correlações entre as respostas Likert e entendê-las em maior profundidade, aplicar a correlação de Spearman, concebida para dados ordinais (as variáveis ordinais têm pelo menos três categorias e as categorias têm uma ordem natural), o que é o caso (Frost, 2021). Os coeficientes de correlação do Spearman variam de -1 a +1:

i. Valores próximos de -1 ou +1 representam relações mais fortes do que valores mais próximos de zero.
ii. Coeficientes fortemente positivos: os valores fortemente concordantes tendem a ocorrer em conjunto.
iii. Coeficientes fortemente negativos: concordo fortemente para uma pergunta é suscetível de coincidir com o Discordo fortemente para outra pergunta.
iv. Coeficientes próximos de zero: Não há relação entre as respostas.

Realizado o estudo de utilizador, os dados devem ser analisados e interpretados de acordo com as indicações fornecidas. As conclusões determinarão se será necessário implementar modificações à interface ou repensar o algoritmo de recomendação.

6. CONCLUSÃO

Esta dissertação de mestrado assumiu como objetivos apresentar uma revisão de literatura sobre os sistemas de recomendação científica; apresentar um plano de testes do sistema de recomendação IViSSEM e da interface, baseado na revisão de literatura; e propor uma métrica quantificadora da reputação científica pessoal, a ser aplicada num contexto de recomendações científicas, e um plano para a validar.

Uma análise aos resultados do trabalho realizado confirma que os objetivos estabelecidos inicialmente foram cumpridos.

Um contributo relevante da dissertação é a perspetiva da revisão de literatura aos sistemas de recomendação: foca-se nos sistemas de recomendação científica, e não nos sistemas de recomendação em geral. A revisão permitiu entender os mecanismos das técnicas de recomendação mais tradicionais – *contente-based filtering, collaborative filtering, hybrid filtering* – e de uma técnica moderna – *graph based*.

Os sistemas de recomendação científica *content-based* utilizam a semelhança entre as características dos itens (por exemplo, título, resumo e palavras-chave) para gerar recomendações; os sistemas de *collaborative-filtering* utilizam a semelhança entre as classificações dos utilizadores para gerar recomendações; os sistemas *graph-based* utilizam gráficos para representar quer os dados dos utilizadores quer os dados dos itens; os sistemas *hybrid* utilizam uma combinação de diferentes técnicas de recomendação para gerar recomendações.

A maior vantagem dos sistemas de recomendação híbridos é as sinergias que formam entre as técnicas de recomendação. Resulta isto na atenuação dos problemas inerentes às restantes técnicas.

Devido à sua mecânica, os sistemas *content*-based sofrem de sobreespecialização, ou seja, recomendam itens demasiado similares entre si, e problemas de semântica relativamente ao conteúdo dos itens. Os sistemas *collaborative-filtering* sofrem de arranque a frio, no qual há uma insuficiência de utilizadores e/ou itens para gerar recomendações, não se adaptam devidamente aos interesses em mudança dos utilizadores, e escassez de dados. Em semelhança, os sistemas *graph-based* também tem como principais desafios o arranque a frio e escassez de dados, em adição à incapacidade de definir associações de alta ordem entre as entidades e às estratégias ineficientes de extração para os meta-padrões.

Refletindo sobre a revisão de literatura, conclui-se que o procedimento de cada um dos três tipos de avaliação – avaliação offline, estudo de utilizador e avaliação offline – segue uma determinada ordem e etapas.

Previamente à avaliação offline, o *dataset* a usar e as métricas de avaliação do sistema de recomendação devem estar definidos, conforme os objetivos estabelecidos. É também necessário selecionar uma estratégia para dividir o *dataset* em conjunto de dados para teste (do sistema) e conjunto de dados para treino (do sistema). Esta divisão e consequentes ações podem ser agilizadas com recurso a *frameworks* e bibliotecas como o Scikit-learn e Surprise. Durante a avaliação, corre-se o código do ou dos algoritmos de recomendação que se quer avaliar, e o código dos algoritmos do estado da arte. Nessa etapa, as métricas são aplicadas e calculadas para cada algoritmo. Subsequentemente, os resultados são comparados para analisar qual deles tem uma melhor performance no geral.

A avaliação online requer que o sistema de recomendação esteja implementado e a ser utilizado por pessoas reais. Elas são realizadas em ambientes controlados, normalmente através dos testes A/B. Primeiro, define-se uma experiência aleatória com duas variantes, A e B. Os utilizadores são depois segmentados com o objetivo de determinar qual dessas variantes tem um impacto mais positivo na experiência do utilizador. Este tipo de avaliação deve ocorrer depois da avaliação offline.

Os estudos de utilizador podem ocorrer em laboratório ou no campo, e dividem-se em estudos quantitativos (medem a satisfação dos utilizadores através de classificações explícitas) e estudos qualitativos (medem a satisfação dos utilizadores através de classificações implícitas).

Nos estudos quantitativos, os utilizadores recebem recomendações geradas por diferentes algoritmos de recomendação e classificam-nas. O algoritmo com a melhor classificação média é considerada a mais eficaz. Nos estudos qualitativos, os participantes são convidados a explorar as recomendações do sistema. Com base nas classificações dadas às recomendações, o sistema retorna outras, que são também perscrutadas pelos participantes. Concluída esta atividade, é distribuído um questionário sobre a qualidade das recomendações. As respostas podem, posteriormente, serem utilizadas para calcular métricas de avaliação.

As escolhas tomadas ao longo da avaliação de um sistema de recomendação são adaptadas consoante as caraterísticas do sistema de recomendação a avaliar e os objetivos específicos dos testes. Com isso em mente, foi prescrito um plano de testes para o sistema de recomendação IViSSEM, que reflete as decisões mais adequadas no contexto do projeto.

Um outro contributo digno de menção é a proposta de uma aceção de reputação científica. Com base na literatura propõe-se a seguinte definição abrange de reputação científica pessoal:

1. É a avaliação agregada que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade em geral relativamente à produtividade e ao impacto dos seus resultados de investigação
2. É o resultado das avaliações que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade em geral relativamente à produtividade e impacto dos seus resultados de investigação, e percurso de vida profissional.
3. É o produto de todo o percurso profissional transposto pelo investigador científico e que define as suas práticas, metodologias e pensamento enquanto impulsionador da ciência.

A reputação é hoje medida com métricas e indicadores tradicionais. Ainda que as *altmetrics* tenham vindo a alterar o panorama, elas não têm tanta adesão em comparação. Por outro lado, a emergência de plataformas científicas como o ResearchGate estão a alterar a perceção dos fatores que englobam a reputação científica.

Querendo contribuir nesse sentido, foi proposta uma métrica quantificadora da reputação para integrar o sistema de recomendação IViSSEM. Ela compreende 20 critérios de 4 categorias distintas: formação e percurso profissional, produtividade, projetos e reconhecimento. O seu contributo passa por levar em consideração todo o percurso profissional do investigador, concedendo perspetivas várias da reputação, e procurar a sua validação junto da comunidade científica.

Como proposta de investigação futura, a partir da dissertação desenvolvida, poderia ser estudada a correlação entre o h-índex, o indicador mais conhecido e aceite, para definir o trabalho de um investigador com os critérios da métrica proposta. O objetivo seria avaliar em que medida o h-índex se alinha com a definição de reputação científica.

Na temática de critérios de reputação, um possível trabalho futuro é a investigação das diferenças entre a perspetiva que os empregadores, entidades financiadoras, governamentais, educativas e similares, possuem sobre a importância dos critérios de reputação propostos para a métrica, e a perspetiva de académicos.

BIBLIOGRAFIA

Ab Latif, R., Mohamed, R., Dahlan, A., & Mat Nor, M. Z. (2016). Using delphi technique: Making sense of consensus in concept mapping structure and multiple choice questions(Mcq). Education in Medicine Journal, 8(3). https://doi.org/10.5959/eimj.v8i3.421

Alarabiat, Alarabiat, A., & Ramos, I. (2019). The delphi method in information sysems research (2004 ofBusiness Research Methods, 17(2). 2017). *Electronic Journal* https://doi.org/10.34190/JBRM.17.2.04

Alfarhood, M., & Cheng, J. (2019). Collaborative attentive autoencoder for scientific article recommendation. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 168–174. https://doi.org/10.1109/ICMLA.2019.00034

Alvelos, F. P. e. (2009). Investigação Operacional: Modelos determinísticos de optimização, métodos e software. Universidade do Minho.

Amami, M., Faiz, R., Stella, F., & Pasi, G. (2017). A graph based approach to scientific paper recommendation. *Proceedings of the International Conference on Web Intelligence*, 777–782. https://doi.org/10.1145/3106426.3106479

Aung, H. H., Zheng, H., Erdt, M., Aw, A. S., Sin, S. J., & Theng, Y. (2019). Investigating familiarity and usage of traditional metrics and altmetrics. *Journal of the Association for Information Science and Technology*, 70(8), 872–887. https://doi.org/10.1002/asi.24162

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific paper recommendation: A survey. IEEE Access, 7, 9324–9339. https://doi.org/10.1109/ACCESS.2018.2890388

Barker, K., Cox, D., Spaapen, J., Van der Meulen, B., Molas Gallert, J., & Sveinsdottir, T. (2011). Social Impact Assessment Methods for research and funding instruments through the study of Productive Interactions between science and society. European Commission.

Barros, M., Moitinho, A., & Couto, F. M. (2019). Using research literature to generate datasets of implicit feedback for recommending scientific items. IEEE Access, 7, 176668–176680. https://doi.org/10.1109/ACCESS.2019.2958002

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0

Belém, F. M., Martins, E. F., Almeida, J. M., & Gonçalves, M. A. (2014). Personalized and object-centered tag recommendation methods for Web 2.0 applications. *Information Processing & Management*, 50(4), 524–553. https://doi.org/10.1016/j.ipm.2014.03.002

Boussaadi, S., Aliane, H., Abdeldjalil, O., Houari, D., & Djoumagh, M. (2020). Recommender systems based on detection community in academic social network. 2020 International Multi-Conference

on: Advanced Technologies" (OCTA), "Organization Knowledge and 1–7. https://doi.org/10.1109/OCTA49274.2020.9151729

Cai, T., Cheng, H., Luo, J., & Zhou, S. (2016). An efficient and simple graph model for scientific article cold start recommendation. Em I. Comyn-Wattiau, K. Tanaka, I.-Y. Song, S. Yamamoto, & M. Saeki (Eds.), Conceptual Modeling (Vol. 9974, pp. 248–259). Springer International Publishing. https://doi.org/10.1007/978-3-319-46397-1_19

Cañamares, R., Castells, P., & Moffat, A. (2020). Offline evaluation options for recommender systems. Information Retrieval Journal, 23(4), 387–410. https://doi.org/10.1007/s10791-020-09371-3

Cervi, C. R., Galante, R., & Oliveira, J. P. M. de. (2013). Comparing the reputation of researchers using a profile model and scientific metrics. 2013 IEEE 16th International Conference on Computational Science and Engineering. https://doi.org/10.1109/cse.2013.61

Chen, Hsinchun (Ed.) (2004): Enhancing Digital Libraries with TechLens+. Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries; Tucson, Arizona, June 7 - 11, 2004. Joint Conference on Digital Libraries; Association for Computing Machinery; Institute of Electrical and Electronics Engineers; ACM/IEEE Joint Conference on Digital Libraries; JCDL 2004. New York, NY: ACM Press.

Cleary, M., Mackey, S., Hunt, G. E., Jackson, D., Thompson, D. R., & Walter, G. (2012). Reputations: A critical yet neglected area of scholarly enquiry: Editorial. *Journal of Advanced Nursing*, 68(10), 2137– 2139. https://doi.org/10.1111/j.1365-2648.2012.06058.x

England, H. F. C. of. (sem data). What is the REF? - REF 2021. Higher Education Funding Council for England. Obtido 15 de Agosto de 2021, de https://www.ref.ac.uk/about/what-is-the-ref/

Ferrara, F., Pudota, N., & Tasso, C. (2011). A keyphrase-based paper recommender system. Em M. Agosti, F. Esposito, C. Meghini, & N. Orio (Eds.), *Digital Libraries and Archives* (Vol. 249, pp. 14–25). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27302-5_2118

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. PLoS ONE, 12(12), e0187394. https://doi.org/10.1371/journal.pone.0187394

Frost,J. (2021, Março 29). Spearman's correlation explained. Statistics by Jim. https://statisticsbyjim.com/basics/spearmans-correlation/

Great Learning Team (2020, setembro 24). What is Cross Validation in Machine learning? Types of Cross Validation. (2021). Consultado em 2 dezembro 2021. Disponível em https://www.mygreatlearning.com/blog/cross-validation/#sh212

Hanafy, R., Makady, S., & El Korany, A. (2018). A social trust metric for scholarly reputation mining. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 61–68. https://doi.org/10.1109/ASONAM.2018.8508251

Hanyurwimfura, Damien; Bo, Liao; Havyarimana, Vincent; Njagi, Dennis; Kagorora, Faustin (2015): An Effective Academic Research Papers Recommendation for Non-profiled Users. IJHIT 8 (3), pág. 255– 272. https://doi.org/10.14257/ijhit.2015.8.3.23.

Harrison, M. (2017, Março 7). eLife collects OrCIDs from authors of accepted papers at proofing. ORCID. https://info.orcid.org/elife-collects-orcids-from-authors-of-accepted-papers-at-proofing/

Haruna K, Akmar Ismail M, Damiasih D, Sutopo J, Herawan T. (2017). A collaborative approach for research paper PLOS ONE 12(10): https://doi.org/10.1371/journal.pone.0184516

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique: Delphi survey technique. *Journal of Advanced Nursing*, 32(4) 1008–1015. https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x

Herman, E. (2018). Scholarly reputation. *FEMS Microbiology Letters*, 365(18). https://doi.org/10.1093/femsle/fny200

Herman, E., & Nicholas, D. (2019). Scholarly reputation building in the digital age: An activity-specific approach. Review article. *El Profesional de la Información*, 28(1). https://doi.org/10.3145/epi.2019.ene.02

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hirsch, J. E., & Buela-Casal, G. (2014). The meaning of the h-index. *International Journal of Clinical and Health Psychology*, 14(2), 161–164. https://doi.org/10.1016/S1697-2600(14)70050-X119

Hui, S., Wei, M., XiaoLiang, Z., JunYan, J., YanBing, L., & ShuJuan, C. (2020). A hybrid paper recommendation method by using heterogeneous graph and metadata. 2020 *International Joint Conference on Neural Networks* (IJCNN), 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206733

IViSSEM. (sem data). Ivissem | information visualization & social scholarly metric. IViSSEM. Obtido 14 de Dezembro de 2021, de http://ivissem.net/

Jamali, H. R., Nicholas, D., & Herman, E. (2016). Scholarly reputation in the digital age and the role of emerging platforms and mechanisms. *Research Evaluation*, 25(1), 37–49. https://doi.org/10.1093/reseval/rvv032

Jiang, Yichen; Jia, Aixia; Feng, Yansong; Zhao, Dongyan (2012): Recommending Academic Papers via Users' Reading Purposes. *Proc. 6th ACMConf. Recommender Syst.* 2012, 2012, pág. 241–244.

Jugovac, M., & Jannach, D. (2017). Interacting with recommenders—Overview and research directions. *ACM Transactions on Interactive Intelligent Systems*, 7(3), 1–46. https://doi.org/10.1145/3001837

Kalachikhin, P. A. (2019). Evaluating researchers based on the criteria of reputation responsibility. *Scientific and Technical Information Processing*, 46(4), 280–287. https://doi.org/10.3103/S0147688219040117

Kanakia, Anshul; Shen, Zhihong; Eide, Darrin; Wang, Kuansan (2019): A Scalable Hybrid Research Paper Recommender System for Microsoft Academic 1, pág. 2893–2899. On-line Disponível em http://arxiv.org/pdf/1905.08880v1.

Kardam, K., Kejriwal, A., Sharma, K., & Kaushal, R. (2016). Ranking scholarly work based on author reputation. *2016 International Conference on Advances in Computing, Communications and Informatics* (ICACCI), 2084–2088. https://doi.org/10.1109/ICACCI.2016.7732358

Kautz, Henry; Selman, Bart; Shah, Mehul (1997): Referral Web. Commun. ACM 40 (3), pág. 63–65. https://doi.org/10.1145/245108.245123

Ma, X., & Wang, R. (2019). Personalized scientific paper recommendation based on heterogeneous graph representation. IEEE Access, 7, 79887–79894. https://doi.org/10.1109/ACCESS.2019.2923293

Maccatrozzo, Valentina; Terstall, Manon; Aroyo, Lora; Schreiber, Guus (03072017): SIRUP. George A. Papadopoulos, Tsvi Kuflik, Fang Chen, Carlos Duarte e Wai-Tat Fu (Ed.): Proceedings of the 22nd International Conference on Intelligent User Interfaces. IUI'17: 22nd International Conference on Intelligent User Interfaces. Limassol Cyprus, 13 03 2017 16 03 2017. New York, NY, USA: ACM, pág. 35–44120

McNee, S. M.; Albert, I.; Cosley, D.; Gopalkrishnan, P.; Lam, S. K.; Rashid, A. et al. (2002): On the recommending of citations for research papers. *Proceedings of the 2002 ACM conference on Computer*, pág. 116–125.

Mohamed Yusoff, A. F., Hashim, A., Muhamad, N., & Wan Hamat, W. N. (2021). Application of fuzzy delphi technique towards designing and developing the elements for the e-pbm pi-poli module. *Asian Journal of University Education*, 17(1), 292. https://doi.org/10.24191/ajue.v17i1.12625

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLoS Medicine*, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Nicholas, D. (2017). New ways of building, showcasing, and measuring scholarly reputation in the digital age. Information Services & Use, 37(1), 1–5. https://doi.org/10.3233/ISU-160792

Nicholas, D., Clark, D., & Herman, E. (2016). Researchgate: Reputation uncovered: researchgate: reputation uncovered. *Learned Publishing*, 29(3), 173–182. https://doi.org/10.1002/leap.1035

Nicholas, D., Herman, E., Xu, J., Boukacem-Zeghmouri, C., Abdullah, A., Watkinson, A., Świgoń, M., & Rodríguez-Bravo, B. (2018). Early career researchers' quest for reputation in the digital age. *Journal of Scholarly Publishing*, 49(4), 375–396. https://doi.org/10.3138/jsp.49.4.01

Osman, N., & Sierra, C. (2016). Reputation in the Academic World. 1578, 1–17.

Pamučar, D., Stević, Ž., & Sremac, S. (2018). A new model for determining weight coefficients of criteria in MCDM models: Full consistency method (FUCOM). *Symmetry*, 10(9), 393. https://doi.org/10.3390/sym10090393

Penfield, T., Baker, M. J., Scoble, R., & Wykes, M. C. (2014). Assessment, evaluations, and definitions of research impact: A review. Research Evaluation, 23(1), 21–32. https://doi.org/10.1093/reseval/rvt021

Philip, Simon; Shola, P.B.; John, Abari Ovye (2014): Application of Content-Based Approach in Research Paper Recommendation System for a Digital Library. (IJACSA) *International Journal of Advanced Computer Science and Application,* 5 (10).

Pu, P., Chen, L. and Hu, R. (2011). A user-centric evaluation framework for recommender systems. *Proceedings of the fifth ACM conference on Recommender systems - RecSys '11*.

REF 2014. (2014). Assessment framework and guidance on submissions. REF 2014. https://www.ref.ac.uk/2014/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf

REF 2021. (2021). Index of revisions to the 'Panel criteria and working methods' (2019/02). Research Excellence Framework. https://www.ref.ac.uk/media/1450/ref-2019_02-panel-criteria-and-working- methods.pdf

Resnick, Paul; Iacovou, Neophytos; Suchak, Mitesh; Bergstrom, Peter; Riedl, John (1994): GroupLens. John B. Smith, F. Don Smith e Thomas W. Malone (Ed.): Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94. the 1994 ACM conference. Chapel Hill, North Carolina, United States, 22-10-1994 - 26-10-1994. New York, New York, USA: ACM Press, pág. 175– 186.

Ricci, Francesco; Rokach, Lior; Shapira, Bracha (2015): Recommender Systems Handbook. Boston, MA: Springer US

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, 15(4), 353–375. https://doi.org/10.1016/S0169-2070(99)00018-7

Ruão, T., & Pessôa, C. (2019). A natureza do fenómeno da reputação científica: O caso dos consórcios universidade-indústria. *Media & Jornalismo*, 19(34), 233–247. https://doi.org/10.14195/2183-5462_34_17

Saat, N. I.; Noah, S. A.; Mohd, M. (2018): Towards serendipity for content-based recommender systems. *International Journal on Advanced Science Engineering and Information Technology 8*, pág. 1762–1769.

Sakib, N., Ahmad, R. B., & Haruna, K. (2020). A collaborative approach toward scientific paper recommendation using citation context. IEEE Access, 8,51246-51255. https://doi.org/10.1109/ACCESS.2020.2980589

Sakib, N., Ahmad, R. B., Ahsan, M., Based, Md. A., Haruna, K., Haider, J., & Gurusamy, S. (2021). A hybrid personalized scientific paper recommendation approach integrating public contextual metadata. *IEEE Access*, 9, 83080–83091. https://doi.org/10.1109/ACCESS.2021.3086964

Shahid, A., Afzal, M. T., Abdar, M., Basiri, M. E., Zhou, X., Yen, N. Y., & Chang, J.-W. (2020). Insights into relevant knowledge extraction techniques: A comprehensive review. *The Journal of Supercomputing*, 76(3), 1695–1733. https://doi.org/10.1007/s11227-019-03009-y

Shani, Guy; Gunawardana, Asela (2011): Evaluating Recommendation Systems.Francesco Ricci, Lior Rokach, Bracha Shapira e Paul B. Kantor (Ed.): Recommender Systems Handbook. Boston, *MA: Springer US*, pág. 257–297.

Sheng Zhang, Weihong Wang, Ford, J., Makedon, F., & Pearlman, J. (2005). Using singular value decomposition approximation for collaborative filtering. *Seventh IEEE International Conference on E-Commerce Technology* (CEC'05), 257–264. https://doi.org/10.1109/ICECT.2005.102

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of- the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734– 749. https://doi.org/10.1109/TKDE.2005.99

Smarandache, F., Ricardo, J. E., Caballero, E. G., Vázquez, Maikel Yelandi Leyva, & Hernández, N. B. (2020). Delphi method for evaluating scientific research proposals in a neutrosophic environment. *Neutrosophic Sets and Systems*, 34(1). https://digitalrepository.unm.edu/nss_journal/vol34/iss1/26

Son, J., & Kim, S. B. (2018). Academic paper recommender system using multilevel simultaneous citation networks. Decision Support Systems, 105, 24–33. https://doi.org/10.1016/j.dss.2017.10.011

Sugiyama, K., Kan, M.-Y. (2015). A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries*, vol. 16, pp. 91-109

Sugiyama, Kazunari, Kan, Min-Yen (2010). Scholarly Paper Recommendation via User's Recent Research Interests. *The 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2010)*, pp.29-38, Gold Coast, Queensland, Australia, June 21-25.

Sugiyama, Kazunari, Kan, Min-Yen (2013). Scholarly Paper Recommendation Datasets. ScholarBank@NUS Repository. [Dataset]. https://doi.org/10.25540/BBCH-QTT8

Sugiyama, Kazunari; Kan, Min-yen (2010): Scholarly Paper Recommendation via User's Recent Research Interests. New York: ACM.

Sun, J., Jiang, Y., Cheng, X., Du, W., Liu, Y., & Ma, J. (2017). A hybrid approach for article recommendation in research social networks. *Journal of Information Science*, 44(5), 696-711. https://doi.org/10.1177/0165551517728449

Sun, J., Ma, J., Liu, Z., & Miao, Y. (2014). Leveraging content and connections for scientific article recommendation in social computing contexts. *The Computer Journal*, 57(9), 1331–1342. https://doi.org/10.1093/comjnl/bxt086

Taneja, A., & Arora, A. (2018). Recommendation research trends: Review, approaches and open issues. *International Journal of Web Engineering and Technology*, 13(2), 123. https://doi.org/10.1504/IJWET.2018.092831

Tang, H., Liu, B., & Qian, J. (2021). Content-based and knowledge graph-based paper recommendation: Exploring user preferences with the knowledge graphs for scientific paper recommendation. *Concurrency and Computation: Practice and Experience*, 33.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner. Consultado em 10 dezembro 2021. Disponível em https://keg.cs.tsinghua.edu.cn/jietang/publications/KDD08-Tang-et-al-ArnetMiner.pdf

Vaidya, J. S. (2005). V-index: A fairer index to quantify an individual's research output capacity. BMJ.

Wang, G., He, X., & Ishuga, C. I. (2018). HAR-SI: A novel hybrid article recommendation approach integrating with social information in scientific social network. Knowledge-Based Systems, 148, 85–99. https://doi.org/10.1016/j.knosys.2018.02.024

Wang, Jun; De Vries, Arjen P.; Reinders, Marcel J. T. (2008): Unified Relevance Models for Rating Prediction in Collaborative Filtering. ACM Transactions on Information Systems 26 (3), pág. 1–40.

Wang, Jun; Vries, Arjen P. de; Reinders, Marcel J.T. (2006): A User-Item Relevance Model for Log-Based Collaborative Filtering. Em: Proc. of ECIR06, pág. 37–48.

Wilsdon, J. (Ed.). (2016). The metric tide: The independent review of the role of metrics in research assessment & management. SAGE.

Yadav, Pratyush; Remala, Nikhila; Pervin, Nargis (2019 - 2019): RecCite: A Hybrid Approach to Recommend Potential Papers. Em: 2019 IEEE International Conference on Big Data (Big Data). 2019 IEEE International Conference on Big Data (Big Data). Los Angeles, CA, USA, 09-12-2019 - 12-12-2019: IEEE, pág. 2956–2964

Zaiontz, C. (sem data). Mann-Whitney Test for Independent Samples. Real Statistics. Obtido 20 de Dezembro de 2021, de https://www.real-statistics.com/non-parametric-tests/mann-whitney-test/

ANEXO I

Universidade do Minho

Conselho de Ética

Comissão de Ética para a Investigação em Ciências Sociais e Humanas

Identificação do documento: CEICSH 124/2021

Relatora: Cristina Maria Moreira Flores

Título do projeto: *Critérios para cálculo da reputação científica*

Equipa de Investigação: Cátia Mendonça (IR), Mestrado em Sistemas de Informação da Universidade do Minho; Prof. Ana Alice Baptista e Prof. Miguel Abrunhosa de Brito (Orientadores), Centro ALGORITMI, Universidade do Minho

PARECER

A Comissão de Ética para a Investigação em Ciências Sociais e Humanas (CEICSH) analisou o processo relativo ao projeto de investigação acima identificado, intitulado *Critérios para cálculo da reputação científica*.
Os documentos apresentados revelam que o projeto obedece aos requisitos exigidos para as boas práticas na investigação com humanos, em conformidade com as normas nacionais e internacionais que regulam a investigação em Ciências Sociais e Humanas.

Face ao exposto, a Comissão de Ética para a Investigação em Ciências Sociais e Humanas (CEICSH) nada tem a opor à realização do projeto nos termos apresentados no Formulário de Identificação e Caracterização do Projeto, que se anexa, emitindo o seu parecer favorável, que foi aprovado por unanimidade pelos seus membros.

Braga, 2 de dezembro de 2021.

O Presidente da CEICSH

(Acílio Estanqueiro Rocha)

Anexo: Formulário de identificação e caracterização do projeto

APÊNDICE I

Critérios para cálculo da reputação científica

A presente investigação enquadra-se no âmbito da dissertação do Mestrado em Sistemas de Informação da Escola de Engenharia da Universidade do Minho, Portugal, sob orientação dos Prof. Doutores Ana Alice Baptista e Miguel Brito.

Pretende-se investigar a perceção de indivíduos de Ciências da Computação em relação à importância de determinados critérios da reputação científica (p.e.: número de publicações). O objetivo é, com os resultados da investigação, calcular os coeficientes dos pesos dos critérios da reputação científica para que estes integrem uma métrica de cálculo da reputação científica.

Por reputação científica entende-se:

1. a avaliação agregada que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade em geral relativamente à produtividade e ao impacto dos seus resultados de investigação.

2. o resultado das avaliações que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade em geral relativamente à produtividade e impacto dos seus resultados de investigação, e percurso de vida profissional.
Exemplos: oferta de financiamento ou de um cargo após avaliação do investigador.

3. o produto de todo o percurso profissional transposto pelo investigador científico e que define as suas práticas, metodologias e pensamento enquanto impulsionador da ciência. Exemplos: publicação de artigos e registo de patentes decorrentes de um trabalho de investigação.

A investigação tem por base um estudo Delphi, composto por este primeiro questionário e um segundo questionário, que depende dos resultados deste. Ele destina-se a indivíduos de centros de investigação sediados em Portugal com experiência profissional em Ciências da Computação.

Este questionário divide-se em duas partes:
Parte I: Informação pessoal - é-lhe pedida informação sobre si e a sua experiência profissional.
Parte II: Comparação dos critérios de reputação científica - é-lhe pedido que compare dois critérios de cada vez e expresse as suas preferências.

A participação é voluntária e pode terminar a qualquer momento sem prejuízo. O questionário é anónimo e confidencial. Tem uma duração de, aproximadamente, Y minutos. As respostas obtidas serão utilizadas apenas para fins de investigação científica e poderão ser consultadas no **repositório institucional da Universidade do Minho**, após publicação da dissertação. Caso pretenda receber informação sobre os resultados do estudo Delphi, envie um email para pg41145@alunos.uminho.pt a expressar a sua intenção.

Agradecemos, desde já, a sua disponibilidade e colaboração no estudo,

Cátia Mendonça

Começar →

Parte I: informação pessoal

Nesta parte, é-lhe pedida informação sobre si e a sua experiência profissional. Para cada questão, selecione a opção que se aplica ao seu caso ou escreva uma resposta.

Género *

A Feminino

B Masculino

C Não-binário

Idade *

A < 26

B 26-35

C 36- 45

D 46-55

E > 55

Anos de experiência profissional na área das Ciências da Computação, Informática, Sistemas de Informação ou afins *

A < 5 anos

B > 5 anos

Ocupação profissional atual *

por exemplo, investigador

País onde exerce a sua atual profissão *

A Portugal

B Outro

Tem experiência direta ou indireta em avaliação de académicos, investigadores ou pares? *

A Sim

B Não

Parte II →

Parte II: Comparação dos critérios de reputação científica

Nesta parte, é-lhe pedido que compare dois critérios de cada vez que dizem respeito à reputação científica e expresse as suas preferências.

O propósito desta tarefa é expressar o nível de importância que acha que o critério *n*-1 tem sobre o critério *n*. Essa preferência é expressada numa **escala de 1 a 9**, onde 1 é a classificação máxima que pode dar a um critério, e 9 a classificação mínima.

Se der a mesma classificação ao critério *n*-1 que ao critério n, então ambos os critérios têm o mesmo nível de importância. Se der, por exemplo, a classificação de 4 ao critério *n*-1 e a classificação 6 ao critério *n*, então o critério *n* é 2 vezes *menos importante* que o critério *n*-1.

Comece por comparar o critério co-autorias com o critério artigos de conferência. **Assuma que artigos de conferência tem a classificação máxima de 1.** De seguida, compare o v-index com co-autorias. Depois, compare peer reviews com v-index e por aí em diante.

Recorda-se que por reputação científica entende-se:

1. a avaliação agregada que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade em geral relativamente à produtividade e ao impacto dos seus resultados de investigação.

2. o resultado das avaliações que um investigador recebe da comunidade científica do domínio em que atua, da comunidade científica em geral e da sociedade em geral relativamente à produtividade e impacto dos seus resultados de investigação, e percurso de vida profissional.

3. o produto de todo o percurso profissional transposto pelo investigador científico e que define as suas práticas, metodologias e pensamento enquanto impulsionador da ciência. Exemplos: publicação de artigos e registo de patentes decorrentes de um trabalho de investigação.

Artigos de conferência

Corresponde a publicações de conferências.

Não dê qualquer classificação, assuma que ela é de 1.

1	2	3	4	5	6	7	8	9

MAIS IMPORTANTE MENOS IMPORTANTE

Co-autorias *

Corresponde a resultados de investigação feitos em colaboração.

Compare com artigos de conferência.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

V-index *

Corresponde ao valor do h-índex ajustado ao ano de publicação. É o h-índex sobre a diferença entre o ano atual e o ano da primeira publicação.

Compare com co-autorias.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Peer reviews *

Corresponde à realização de revisão de pares.

Compare com v-index.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Financiamento *

Corresponde ao financiamento recebido por entidades públicas ou privadas para a investigação.

Compare com peer reviews.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Livros *

Corresponde à publicação de livros.

Compare com financiamento.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Artigos de revista *

Corresponde a publicações em revistas.

Compare com livros.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Propriedade intelectual *

Corresponde a patentes registadas ou pedidos provisórios de registo de patentes.

Compare com artigos de revista.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Educação *

Corresponde à afiliação do investigador com universidades e respetivos graus académicos.

Compare com propriedade intelectual.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Emprego *

Corresponde à atividade profissional atual e passadas do investigador.

Compare com educação.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Capítulos de livros *

Corresponde à escrita de um ou mais capítulo de livros publicados em colaboração.

Compare com emprego.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Livros editados *

Corresponde a publicações de livros editados parcial ou totalmente.

Compare com capítulos de livros.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Manuais *

Corresponde a publicações de manuais escolares ou académicos.

Compare com livros editados.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Relatórios *

Corresponde à publicação de relatórios.

Compare com manuais.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Prémios e outras distinções *

Corresponde a prémios e distinções recebidos por entidades ao longo da carreira.

Compare com relatórios.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Qualificações *

Corresponde à prova da aquisição de competências como certificados e formação profissional.

Compare com prémios e outras distinções.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Serviços *

Corresponde a atividades realizadas ao serviço de uma organização.

Compare com qualificações.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Recursos de investigação *

Corresponde a recursos oferecidos para trabalhos de investigação.

Compare com serviços.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Convites para cargos de trabalho *

Corresponde a ofertas de trabalho aceites.

Compare com recursos e investigação.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Associações membro *

Corresponde à afiliação do investigador com organizações ou sociedades enquanto membro.

Compare com convites para cargos de trabalho.

1	2	3	4	5	6	7	8	9

MENOS IMPORTANTE MAIS IMPORTANTE

Submeter →

APÊNDICE II

Formulário de consentimento informado, livre e esclarecido

Título do projeto: IViSSEM – 6.849,32 Journal Articles Everyday: Visualize or Perish!

Pessoa responsável pelo projeto: Ana Alice Rodrigues Pereira Baptista

Instituição responsável: Universidade do Minho

Informação geral

(a preencher)

Qual é o objetivo do estudo?

O estudo visa compreender (1) de que forma os utilizadores da plataforma IViSSEM interagem com a mesma, (2) se a interface permite uma navegação satisfatória para o participante, e (3) se a plataforma fornece recomendações relevantes de research outputs ao participante.

Qual é o meu papel no estudo?

O participante será convidado a efetuar um conjunto de atividades na plataforma IViSSEM, relacionadas com os objetivos do estudo.

Qual a duração esperada da minha participação?

A sua participação será feita em três dias distintos, numa sessão que tem a duração de, no máximo, (preencher).

A minha participação é voluntária?

A participação na investigação é voluntária. Pode desistir a qualquer momento ou recusar-se a participar sem qualquer tipo de consequência.

Quais os procedimentos do estudo em que vou participar?

Antes do início do estudo será feita uma introdução ao projeto e ao estudo, incluindo a contextualização da sua participação. Ser-lhe-ão dadas instruções sobre as tarefas que terá de realizar. Por fim, após o estudo ser-lhe-á pedido que responda a um questionário sobre a sua experiência.

Que dados serão recolhidos?

A realização do estudo será gravada em áudio/vídeo digital e o participante será fotografado.

O que irá acontecer aos meus dados?

Na eventual publicação de artigos ou na participação em conferências, a sua identidade não será revelada ao público e nenhuma conclusão sobre a sua identidade a partir de dados publicados poderá ser tirada. Os dados por si fornecidos só serão utilizados no projeto IViSSEM e para fins científicos, sem que seja identificado(a).

O que acontecerá aos meus dados após a conclusão do projeto?

Os dados relativos à sua identidade serão apagados após o término do projeto IViSSEM ou armazenados de forma anónima.

Como me posso informar acerca dos resultados do estudo do projeto?

Os resultados serão publicados no formato de artigos científicos e/ou apresentados em conferências, abertos a toda a comunidade. Em alternativa, pode contactar-nos e inquirir sobre os resultados do estudo do projeto.

Quem deverei contactar em caso de dúvidas sobre o projeto e o meu envolvimento no projeto?

Para qualquer dúvida relacionada, contactar (preencher) através (preencher).

Assinatura do consentimento informado, livre e esclarecido

_____

Nome do participante

_____

Data

APÊNDICE III

Número da tarefa	Descrição da tarefa
1.1	Crie uma conta.
1.2	Preencha o seu perfil com os seus domínicos de conhecimento.
2.1	Consulte os detalhes (metadados) de um artigo científico.
2.2.1	Rejeite uma recomendação ao interagir com o gráfico na página inicial.
2.2.2*	Aceite uma recomendação.
2.3	Adicione uma recomendação à sua biblioteca pessoal.
2.3.1	Abra um artigo científico como se o fosse ler.
2.3.2	Recomende o artigo científico que abriu.
2.4.2.2	Rejeite uma recomendação através do menu/painel da página inicial.
3.1	Na visualização *time travel* filtre as recomendações por data.
4.1	Adicione um artigo científico à plataforma.
5.1	Procure por um recomendador e explore o perfil deste.
5.2	Dê *follow* a um recomendador.
5.3	Adicione uma recomendação de um recomendador – que esteja a seguir ou outro.
6.1	Na página inicial, para cada umas das recomendações, aceite ou rejeite a recomendaçãoconforme o seu entender.
6.1.2	Na página *My Library* aparece-lhe as recomendações aceites por si. Tendo lido, pelo menos, o resumo de cada artigo científico, classifique-os clicando nos *smiles*.

APÊNDICE IV

Questionário ResQue

Os utilizadores devem indicar as suas respostas para cada uma das perguntas utilizando a Escala de Likert de 1 a 5, em que:

1 corresponde a “Discordo totalmente”,

2 corresponde a "Discordo",

3 corresponde a "Nem concordo nem discordo",

4 corresponde a "Concordo",

5 corresponde “Concordo totalmente”.

A) Qualidade dos Resultados de investigação recomendados

1. Os Resultados de investigação recomendados são relevantes para os meus interesses e domínio científico.

2. Os Resultados de investigação que me foram recomendados são novos para mim e interessantes.

3. O sistema de recomendação ajudou-me a descobrir novos Resultados de investigação.

4. O sistema de recomendação não me permitiu encontrar novos Resultados de investigação.

5. Os Resultados de investigação recomendados são diversificados

6. Os Resultados de investigação recomendados são semelhantes entre si.

7. Os Resultados de investigação recomendados levaram em consideração o meu perfil.

B) Adequação da interação

8. A plataforma fornece uma forma adequada de expressar as minhas preferências.

9. A plataforma proporciona uma forma adequada de rever as minhas preferências.

10. A plataforma explica porque é que os produtos me são recomendados.

C) Adequação da interface

11. A interface da plataforma apresenta informações suficientes para me conseguir orientar.

12. A informação dada sobre um RO recomendado é suficiente para eu decidir se ele é relevante ou não.

13. As legendas/informação da interface da plataforma são claras e adequadas.

14. Consegui completar rapidamente as tarefas e cenários utilizando este sistema.

15. Se utilizasse esta interface novamente, eu provavelmente realizaria as tarefas mais rapidamente.

16. Foi fácil de aprender a utilizar este sistema.

17. O layout da interface é atrativo e adequado.

18. Este sistema tem todas as funções e capacidades que procuro num sistema de recomendação.

D) Perceção de facilidade de utilização

19. Familiarizei-me muito rapidamente com o sistema de recomendação.

20. Encontrei facilmente os Resultados de investigação recomendados.

21. A procura dos Resultados de investigação recomendados exigiu demasiado esforço.

22. Achei fácil indicar ao sistema as minhas preferências.

23. É fácil treinar o sistema para atualizar as minhas preferências.

24. É fácil dizer ao sistema se gosto/não gosto do RO recomendado.

25. Na minha opinião, é fácil obter um novo conjunto de recomendações.

26. Encontrar um RO de interesse com a ajuda do recomendador é fácil.

27. Penso que me poderia tornar produtivo rapidamente utilizando este sistema.

E) Perceção de utilidade

28. De modo geral, estou satisfeito com sistema de recomendação.

29. As recomendações ajudaram-me efetivamente a encontrar Resultados de investigação relevantes.

30. Sinto-me apoiado para encontrar Resultados de investigação relevantes com a ajuda do recomendador.

F) Transparência

31. Compreendi porque é que os Resultados de investigação me foram recomendados.

32. O sistema ajuda-me a compreender porque é que os Resultados de investigação me foram recomendados.

G) Atitudes

33. Confio nos Resultados de investigação que me foram recomendados, uma vez que eram coerentes com o meu perfil e interesses.

34. Estou confiante de que irei gostar dos Resultados de investigação que me foram recomendados.

H) Intenções Comportamentais

35. Voltaria a utilizar este sistema de recomendação.

36. Dada a oportunidade, utilizaria este sistema de recomendação com frequência.

37. Recomendaria aos meus amigos/colegas esta plataforma.

38. Utilizaria os Resultados de investigação recomendados na minha investigação.

APÊNDICE V

Questinário ResQue Participantes

Identificação do participante:

Data:

Numa escala de 1 a 5, onde

1 corresponde a “Discordo totalmente”,

2 corresponde a "Discordo",

3 corresponde a "Nem concordo nem discordo",

4 corresponde a "Concordo",

5 corresponde “Concordo totalmente”,

classifique as seguintes afirmações:

A) Qualidade dos resultados de investigação recomendados	
1. Os resultados de investigação recomendados são relevantes para os meusinteresses e domínio científico.	
2. Os resultados de investigação que me foram recomendados são novos paramim e interessantes.	
3. O sistema de recomendação ajudou-me a descobrir novos resultados deinvestigação.	
4. O sistema de recomendação não me permitiu encontrar novos resultados deinvestigação.	
5. Os resultados de investigação recomendados são diversificados	
6. Os resultados de investigação recomendados são semelhantes entre si.	
7. Os resultados de investigação recomendados levaram em consideração o meuperfil.	
B) Adequação da interação	
8. A plataforma fornece uma forma adequada de expressar as minhaspreferências.	
9. A plataforma proporciona uma forma adequada de rever as minhaspreferências.	
10. A plataforma explica porque é que os produtos me são recomendados.	
C) Adequação da interface	
11. A interface da plataforma apresenta informações suficientes para me conseguirorientar.	

12. A informação dada sobre um RO recomendado é suficiente para eu decidir se ele é relevante ou não.	
13. As legendas/informação da interface da plataforma são claras e adequadas.	
14. Consegui completar rapidamente as tarefas e cenários utilizando este sistema.	
15. Se utilizasse esta interface novamente, eu provavelmente realizaria as tarefas mais rapidamente.	
16. Foi fácil de aprender a utilizar este sistema.	
17. O layout da interface é atrativo e adequado.	
18. Este sistema tem todas as funções e capacidades que procuro num sistemade recomendação.	
D) Perceção de facilidade de utilização	
19. Familiarizei-me muito rapidamente com o sistema de recomendação.	
20. Encontrei facilmente os resultados de investigação recomendados.	
21. A procura dos resultados de investigação recomendados exigiu demasiadoesforço.	
22. Achei fácil indicar ao sistema as minhas preferências.	
23. É fácil treinar o sistema para atualizar as minhas preferências.	
24. É fácil dizer ao sistema se gosto/não gosto dos resultados de investigaçãorecomendados.	
25. Na minha opinião, é fácil obter um novo conjunto de recomendações.	
26. Encontrar um RO de interesse com a ajuda do recomendador é fácil.	
27. Penso que me poderia tornar produtivo rapidamente utilizando este sistema.	
E) Perceção de utilidade	
28. De modo geral, estou satisfeito com sistema de recomendação.	
29. As recomendações ajudaram-me efetivamente a encontrar resultados deinvestigação relevantes.	
30. Sinto-me apoiado para encontrar resultados de investigação relevantes com aajuda do recomendador.	
F) Transparência	
31. Compreendi porque é que os resultados de investigação me foramrecomendados.	
32. O sistema ajuda-me a compreender porque é que os resultados de investigaçãome foram recomendados.	
G) Atitudes	
33. Confio nos resultados de investigação que me foram recomendados, uma vezque eram coerentes com o meu perfil e	

interesses.	
34. Estou confiante de que irei gostar dos resultados de investigação que me foramrecomendados.	
H) Intenções Comportamentais	
35. Voltaria a utilizar este sistema de recomendação.	
36. Dada a oportunidade, utilizaria este sistema de recomendação com frequência.	
37. Recomendaria aos meus amigos/colegas esta plataforma.	
38. Utilizaria os resultados de investigação recomendados na minha investigação.	