Como citar este artículo: M. S. Stein, C. Cechinel and V. F. C. Ramos, "Quantitative Analysis of Users' Agreement on Open Educational Resources Quality Inside Repositories," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 18, no. 1, pp. 2-9, Feb. 2023, doi: 10.1109/RITA.2023.3250446.

Análise Quantitativa da Concordância de Avaliadores de Recursos Educacionais Digitais dentro de Repositórios

Mayara Sousa Stein*, Cristian Cechinel*, and Vinicius Ramos*

***Abstract*—Quality assessment inside learning object repositories is normally performed by the community of users that share interest and rate the same resources. At the same time, this strategy is largely disseminated in the most known repositories. In addition, the final presentation of the overall quality of the resources is normally restricted to the average rating given by the community, thus, hiding the internal distribution of the ratings and the characteristics of the users involved in the evaluation process. The present paper analyzes to which extent different raters tend to agree about the quality of the resources inside the Merlot repository. For that, data were collected from the repository and calculated the Intra-Class Correlation coefficient for 102 pairs of evaluators, as well as the Spearman correlation among the average ratings of a given resource by evaluators coming from the same categories of disciplines. Results point out a high concentration of poor agreement between raters (75% to 85% of the pairs of raters tended to disagree), and no correlation among the average ratings of the resources from the different disciplines. Based on these findings, the authors suggest improvements to the repository interface better presenting the overall quality of the resources.**

***Index Terms*—User-generated content, Educational technology, Multimedia databases, Online services, Open Educational Resources**

I. Introdução

OBJETOS de Aprendizagem (OA) têm um papel importante no processo de ensino-aprendizagem na era digital. Para o professor, um destes papéis é a possibilidade de melhoria das práticas para elaboração dos materiais de ensino, como uma forma de unificar o seu formato e sua estrutura. Muitas vezes, os professores elaboram materiais que nunca mais são utilizados ou não são criados com a possibilidade de reutilização. Ainda, criadores podem elaborar materiais já existentes ou que se encontram em formatos de difícil edição. Para os estudantes, tal importância varia desde a facilidade de encontrar recursos digitais para a sua disciplina/conteúdo, até a utilização de recursos completos que envolvem a apresentação de conteúdo e, também, a avaliação do conhecimento.

Visando potencializar a sua utilização, estes recursos são, normalmente, armazenados, compartilhados, mantidos e utilizados em plataformas digitais chamadas de Repositórios de Objetos de Aprendizagem (ROA). Em virtude do crescimento do número de OA, e como uma forma de estimular o seu uso, além de qualificar e facilitar sua busca, é essencial que os ROA adotem estratégias para garantir a qualidade desses objetos.

Cechinel [1] afirma que a garantia da qualidade de recursos em repositórios é uma tarefa complexa, pois ela incorpora diferentes aspectos e dimensões, como a qualidade do conteúdo e a usabilidade num âmbito amplo. Além disso, pode-se afirmar que a excelência dos recursos é de fundamental importância para o próprio repositório, sendo que estes repositórios aproveitam a colaboração e as interações dos seus usuários, oferecendo-lhes a possibilidade de avaliar e opinar sobre os diversos objetos neles armazenados [2]. A qualidade dos OA é importante para os repositórios uma vez que os algoritmos de ranqueamento desses sistemas usam as informações sobre as avaliações da comunidade para posicioná-los após uma busca realizada pelos usuários, ou, ainda, podem ser usadas pelos algoritmos de recomendação (quando estes estão implementados) e, muitas vezes, também são usadas como referência para as buscas dos usuários de conteúdo específico [3].

Alguns repositórios utilizam instrumentos específicos , sendo que os recursos são avaliados pela comunidade de usuários [2][4]. Exemplos de repositórios que utilizam esse tipo de avaliação são: Common Sense Education[1], o Merlot[2] e a Plataforma MEC de recursos educacionais digitais[3].

Quando se trata de avaliação, os principais ROA existentes as representam como uma pontuação final única, resultante da média das notas atribuídas pelos seus usuários. Isso esconde a distribuição das notas dadas, além de impossibilitar ao usuário conhecer a opinião de grupos de usuários similares. Por exemplo, um determinado OA para a área de Artes pode ter sido bem avaliado por usuários da mesma área, porém mal avaliado por usuários de outras áreas. Estas nuances, referentes às avaliações dos diferentes grupos, simplesmente desaparecem na apresentação da qualidade dos mesmos dentro dos ROA. Assim, usando o exemplo acima, um usuário da área de Artes poderia não encontrar facilmente o OA bem avaliado na sua área porque outras áreas fizeram a sua pontuação baixar e, portanto, ser pessimamente ranqueado na apresentação dos resultados. Nesta linha, pesquisas anteriores demonstraram que existe considerável discordância com relação à qualidade

Mayara Sousa Stein é mestre pelo Programa de Pós-graduação em Tecnologias da Informação e Comunicação (PPGTIC) da Universidade Federal de Santa Catarina, Campus Araranguá, Unidade Mato Alto, Rua Pedro João Pereira, 150, Mato Alto – Araranguá – SC, CEP 88.905-120 (e-mail: stein.mayara@gmail.com).

Cristian Cechinel e Vinicius Ramos são professores do PPGTIC da UFSC (e-mails: contato@cristiancechinel.pro.br, v.ramos@ufsc.br)

Manuscript received May 05, 2020; revised July 06, 2020.

[1]https://www.commonsense.org/education. Acessado em 13/06/2020.

[2]https://www.merlot.org. Acessado em 13/06/2020.

[3]https://plataformaintegrada.mec.gov.br. Acessado em 13/06/2020.

dos recursos entre a comunidade de usuários do Merlot e a comunidade de especialistas avaliadores [5], [6].

Neste contexto, este trabalho apresenta um estudo diretamente relacionado às avaliações dos OA do repositório Merlot que foram realizadas pela sua comunidade de usuários, com uma abordagem de avaliação aberta e descentralizada [7]. Este estudo baseia-se, portanto, em uma análise quantitativa da concordância entre pares de usuários avaliadores com relação à qualidade dos OA avaliados por eles. Ademais, apresentam-se sugestões para a melhoria da qualidade da interface de apresentação da pontuação final, buscando mostrar, com maiores detalhes, a variante das avaliações dos usuários e seus atributos relacionados (área de atuação, especialidade, entre outras).

A partir destas considerações, o presente trabalho se propõe a responder às seguintes perguntas de pesquisa:

- **P1**: Qual a concordância geral entre pares de avaliadores usuários de recursos educacionais digitais?
- **P2:** Qual o nível de concordância entre pares de avaliadores usuários considerando suas preferências por determinadas disciplinas e suas áreas de atuação profissional?
- **P3**: É possível identificar alguma associação significativa entre as notas dadas pelos avaliadores usuários e as disciplinas de preferência dos avaliadores e dos recursos?

O restante do artigo está estruturado da seguinte maneira. A Seção II discute brevemente algumas medidas de consistência existentes. A Seção III descreve os dados utilizados e a metodologia adotada no estudo e a Seção IV descreve os principais resultados encontrados. Finalmente, a Seção V discute as descobertas do trabalho e algumas aplicações, e a Seção VI apresenta as considerações finais e as possibilidades de trabalhos futuros.

II. Concordância e confiabilidade entre avaliadores

Existem basicamente dois termos principais utilizados para trabalhar com o cálculo da consistência entre as pontuações dadas por diferentes avaliadores (às vezes chamados de juízes), sendo eles a *concordância* e a *confiabilidade*. Muitas vezes estes termos são utilizados de maneira intercambiável e generalizada, no entanto, confiabilidade e concordância de avaliadores possuem algumas diferenças importantes de serem mencionadas [8], [9].

Alguns autores definem *confiabilidade* como a medida de consistência entre avaliadores, seja na classificação ou na posição relativa das classificações de desempenho/notas fornecidas [8], [10]. Outros autores definem *concordância* como a medida em que dois ou mais avaliadores, que usam a mesma escala de classificação e avaliam os mesmos materiais, fornecem as mesmas notas, ou seja, a concordância mede a frequência com que os avaliadores atribuem exatamente a mesma classificação para os recursos avaliados [11], [12].

O coeficiente Kappa de Cohen é uma medida de concordância muito utilizada na literatura. Esta medida é usada para variáveis qualitativas, inclusive binárias (certo e errado, bom e ruim, etc), sendo que sua forma tradicional permite a comparação entre as classificações de dois avaliadores em um mesmo objeto de observação. Esta medida, ao ser calculada, considera a concordância ao acaso (é definida a proporção de concordância devido ao acaso Pa), o coeficiente pode então ser definido como a proporção de concordância entre os juízes (Po) menos a proporção de acaso ($Po - Pa/1 - Pa$). O coeficiente de concordância possui o limite máximo de 1, representando a concordância perfeita entre os juízes, quanto mais próximo de 0 estiver o coeficiente indica que a concordância entre os juízes ocorre por acaso [13]. A correlação intraclasse (ICC, por seu acrônimo em inglês) é uma outra medida de concordância, mas que também pode ser utilizada para medir a confiabilidade. Esta medida permite que o cálculo de concordância seja feito entre avaliadores para dados quantitativos. Por outro lado, a comparação é feita a partir da variabilidade de diferentes classificações do mesmo objeto com a variação total em todas as classificações dos objetos [14]. Existem ainda outros métodos para o cálculo da consistência entre avaliadores, como o Kappa Ponderado (concordância) e o Alpha de Krippendorf (confiabilidade e concordância) [9].

No presente trabalho, utilizou-se o ICC como medida de concordância das avaliações entre pares de avaliadores. Com o objetivo de evitar diferentes análises dos resultados, utiliza-se a interpretação de Fleiss, Levin e Paik [15] , a qual sugere que:

- ICC < 0,4: indicativo de uma concordância baixa ou pobre.
- 0,4 <= ICC < 0,75: indicativo de uma concordância moderada ou satisfatória.
- ICC >= 0,75: indicativo de concordância excelente entre avaliadores.

Para escolher a fórmula do ICC, foi utilizado o método de decisão para correlação intraclasse apresentado no trabalho de Koo e Li [14]. Na literatura, costuma-se dividir o ICC em três modelos básicos [16]: **efeito aleatório de uma via/unidirecional** (*one-way random*) - cada objeto é avaliado por um conjunto distinto de avaliadores, escolhidos aleatoriamente no conjunto de dados; **efeito aleatório de duas vias/bidirecional** (*two-way random*): - os avaliadores são selecionados aleatoriamente de acordo com suas características semelhantes; e **efeito misto bidirecional** (*two-way mixed*) - este modelo é utilizado quando os avaliadores selecionados são os únicos avaliadores de interesse.

O ICC calculado considera cada par de avaliadores que possuem conjuntos de recursos comumente avaliados, e a escolha desses pares não é aleatória (e sim um consenso), portanto, optou-se por utilizar o ICC de efeito misto bidirecional. Para o tipo, utiliza-se o valor médio de K avaliadores, sendo possível utilizar o modelo de decisão ICC (3, K), que refere-se à fórmula modelo efeito misto bidirecional. Neste caso, o número 3 da fórmula indica que o conjunto de dados analisados é classificado pelos mesmos avaliadores e o K representa o número de avaliadores a serem considerados.

A fórmula computacional do ICC (3,K) utiliza as seguintes equações: $MS_B = \frac{k}{n-1}\sum_{j=1}^{n}(\bar{y}._j - \bar{y}..)^2$ onde $\bar{y}.j$ é a média das avaliações do avaliador j e $\bar{y}..$ é a média de todas as avaliações. E $MS_E = \frac{k}{(n-1)(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}i. - \bar{y}.j +$

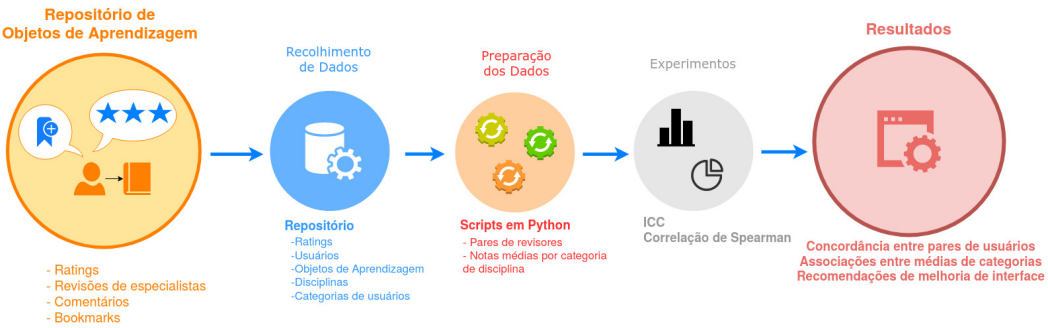

Fig. 1. Esboço geral da metodologia

$\bar{y}..)^2$, na qual, $\bar{y}i.$ é a média das avaliações de i documentos e y_{ij} é o valor da avaliação i do avaliador j. Desta forma, segue a fórmula computacional utilizada $\frac{MS_B - MS_E}{MS_B}$. A variável MS_E refere-se ao quadrado médio do erro, MS_B ao quadrado médio entre os sujeitos e K é a quantidade de avaliadores divida pelas medições [14].

III. Materiais e Métodos

Os dados utilizados neste estudo foram extraídos do repositório Merlot (ano 2015). O Merlot[4] é uma iniciativa internacional que busca fornecer um ambiente gratuito, de fácil acesso, que conecta diferentes tipos de pessoas (professores, funcionários, alunos, entre outros) a recursos educacionais digitais [17]. O repositório é formado por mais de 90.000 materiais pertencentes a 19 categorias diferentes (Artes, Negócios, Tecnologias, entre outras). Os materiais inseridos no Merlot são fornecidos pelos membros cadastrados na plataforma. Todos os serviços oferecidos pela plataforma são totalmente gratuitos. O Merlot é classificado como um referatório [18] devido a forma de armazenamento dos recursos educacionais digitais, uma vez que a plataforma mantém apenas os metadados de cada recurso junto com um link para acesso externo que redireciona o usuário para o destino original do mesmo.

No Merlot, os recursos educacionais digitais são avaliados por duas comunidades distintas, os usuários da comunidade e os especialistas (peer-reviewers). A avaliação realizada pela comunidade de usuários não utiliza critérios de qualidade padronizados e consiste em uma escala de pontuação que vai de 1 a 5, além de comentários dos usuários com relação a pontuação marcada por eles. A avaliação dos especialistas é organizada pelos conselhos editoriais, que distribuem os recursos para serem avaliados pelos membros especialistas das áreas específicas dos recursos. Essa avaliação é realizada por meio de pontuações de 1 a 5 e comentários sobre três critérios principais de qualidade: qualidade de conteúdo, efetividade, e facilidade de uso. Esses membros especialistas são previamente treinados para que compreender os critérios e processos de avaliação. Somente os recursos considerados mais importantes pelo Merlot são destinados a avaliação por parte dos especialistas. Trabalhos anteriores já identificaram uma divergência entre as notas fornecidas pela comunidade de especialistas e pela comunidade de usuários do Merlot [19].

[4]https://www.merlot.org/

Por meio de um crawler[20], foram recolhidas um total de 15.886 avaliações de recursos educacionais, sendo 11.823 avaliações realizadas pela comunidade de usuários e 4.064 avaliações realizadas pelos especialistas, para os mais de 62.000 recursos cadastrados (até a data de recolhimento) nas diferentes áreas do conhecimento (ciência e tecnologia, humanas, negócios, matemática, artes, educação).

Foram realizadas duas análises sobre os recursos:

1) Experimento 1 - concordância entre os pares de avaliadores usuários, e
2) Experimento 2 - correlação entre as médias das notas de cada categoria de disciplina.

Para cada análise foi necessário um pré-processamento específico de dados por meio de um script Python. No Merlot, apenas as avaliações fornecidas pelos usuários possuem identificação. Considerando isso, o presente trabalho considerou apenas as avaliações realizadas pela comunidade de usuários, uma vez que essas avaliações permitiam a identificação dos usuários que avaliaram um mesmo recurso e a consequente formação dos pares de avaliadores. A Fig. 1 mostra o esboço geral da metodologia seguida.

A. Preparação dos dados para o Experimento 1 - Concordância entre os pares de usuários

A base de dados possui o id e o nome do avaliador usuário, as notas fornecidas pelo mesmo, a disciplina relacionada à sua profissão (ciência e tecnologia, humanas, negócios, etc) e sua atuação profissional (estudante, professor, bibliotecário, etc). Recursos com menos de 2 avaliações foram excluídos da análise, resultando em 1.452 avaliações de usuários de 11.823 existentes. Desse conjunto, avaliadores que pontuaram simultaneamente um mesmo recurso foram separados em pares, resultando em 179 pares de usuários, que avaliaram um total de 925 recursos de aprendizagem. A Fig. 2 apresenta a distribuição da quantidade de recursos avaliados simultaneamente por pares (duplas) de avaliadores. Os pares de usuários que realizaram 3 ou menos avaliações simultâneas (39 pares) foram excluídos da análise por ser uma quantidade pequena de avaliações, o que pode ocasionar uma concordância ao acaso [14], resultando em 140 pares de usuários. Por último, foram retirados outros 38 pares de usuários, em que as avaliações não possuíam a nota quantitativa (apenas o comentário). Os 102 pares de usuários resultantes foram utilizados na análise de concordância entre os pares de avaliadores usuários.

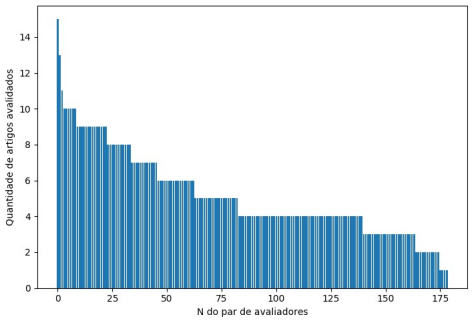

Fig. 2. Quantidade de materiais avaliados por pares de usuários

TABLE I
CATEGORIAS EXISTENTES E SUA UTILIZAÇÃO NOS EXPERIMENTOS

Categoria	Explicação	Exemplo\Comentário
Comunidade de usuários	Qualquer usuário registrado na plataforma	As pontuações dessa comunidade foram utilizadas nos experimentos)
Comunidade de especialistas	Especialistas registrados pelos editores das diferentes áreas de conhecimento do Merlot	As pontuações dessa comunidade **não** foram utilizadas nos experimentos)
Disciplina preferencial	Disciplina de preferência que o usuário registrou como sendo de sua maior especialidade ou interesse	Ciência e tecnologia, ciências sociais, humanas, negócios, matemática, artes e educação)
Atuação profissional	Profissão no setor educacional que o usuário ocupa	Professor (universitário ou não), estudante, bibliotecário, funcionário, consultor)

Todos os pré-processamentos e cálculos iniciais foram realizados através de *scripts* em Python. Com os pares de avaliadores separados com suas respectivas avaliações, calculou-se o coeficiente ICC para cada par de usuário avaliador. Após o cálculo, realizaram-se as análises das concordâncias de acordo com a área de atuação preferencial (categoria de disciplina) e a classe de atuação profissional . Os resultados são apresentados nas Seções IV-A e IV-B, respectivamente.

De maneira geral, as áreas de atuação preferencial são idênticas às categorias de disciplinas em que os OA são cadastrados (ciência e tecnologia, ciências sociais, humanas, negócios, matemática, artes e educação). Essas categorias primárias, por sua vez, são subdivididas em subcategorias secundárias (por exemplo, ciência e tecnologia é subdividida em Biologia, Física, Química, Tecnologia da Informação, entre outras). Por outro lado, as classes de atuação profissional representam a atividade desenvolvida pelo usuário (professor não universitário - *teacher*, estudante, profissional universitário - professor universitário, pesquisador, assistente técnico administrativo em educação, bibliotecário, entre outros). A tabela I descreve essas categorias utilizadas. Além do percentual do ICC para cada uma das categorias de disciplina e de área profissional, sempre que possível, também foi calculado o qui-quadrado para avaliar se estas categorias estão associadas com o nível de concordância sobre a qualidade dos recursos.

B. Preparação dos dados para o Experimento 2 - Correlação entre as médias das notas de cada categoria de disciplina

Neste experimento, foi avaliada a existência de associação entre as médias das avaliações dadas pelos avaliadores de cada uma das categorias de disciplinas. Desta maneira, cada linha da tabela resultante tem o recurso avaliado e a média das notas de cada uma das 8 áreas de atuação preferencial dos avaliadores. No total foram consideradas as avaliações de 2.220 recursos.

Com a tabela resultante do pré-processamento dos dados, foram relacionadas as médias obtidas em cada categoria e aplicado um teste de correlação entre as colunas. Após verificar que as amostras não seguem uma distribuição normal, utilizou-se a correlação de *Spearman* para avaliar a existência de correlação entre as médias das diferentes disciplinas. Ou seja, para verificar a concordância entre as médias das notas dadas pelos avaliadores usuários e a categoria dos recursos.

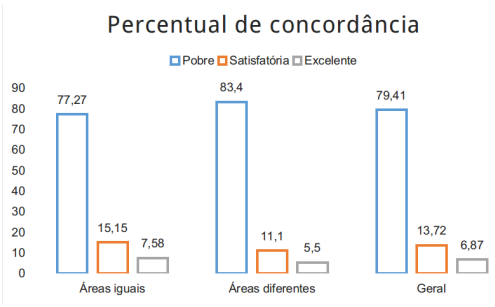

Fig. 3. Concordância geral por disciplinas preferenciais iguais e diferentes

IV. RESULTADOS

A. Experimento 1A - Concordância entre pares de avaliadores usuários considerando disciplinas de atuação

A Fig. 3 apresenta os percentuais dos níveis de concordância dos pares de avaliadores usuários considerando as situações em que ambos pertencem a uma mesma disciplina preferencial (disciplina primária) e para quando pertencem a disciplinas preferenciais distintas, além do nível de concordância geral sem considerar a disciplina preferencial.

Como pode ser visto na Fig. 3, 79,41% dos pares de avaliadores possuem um nível de concordância pobre (ICC $<$0,4), enquanto 13,72% possuem um nível de concordância satisfatória (0,4 $<=$ ICC $<$ 0,7) e apenas 6,87% possuem um nível de concordância excelente (ICC $>=$ 0,7). O percentual de concordâncias satisfatória e excelente entre os pares de avaliadores que pertencem a uma mesma disciplina preferencial é levemente maior do que os percentuais de concordância entre os pares de avaliadores com disciplinas preferenciais distintas. Enquanto no primeiro caso os percentuais de concordância satisfatória e excelente são de 15,15% e 7,58%, respectivamente, no segundo cenário estes percentuais caem para 11,1% e 5,5%.

Foi realizada também uma análise de associação entre a concordância observada nos diferentes grupos (disciplina preferencial igual versus disciplina preferencial diferente). De acordo com [21], o método qui-quadrado permite testar a significância da associação entre duas variáveis qualitativas. As hipóteses do teste foram formuladas da seguinte maneira:

- H_0: Os níveis de concordância são independentes da disciplina preferencial do avaliador
- H_1: Existe associação entre os níveis de concordância e a disciplina preferencial do avaliador

Para a execução do teste, agrupamos os níveis de concordância satisfatório e excelente em um único grupo. O teste qui-quadrado (χ^2) mostrou que as diferenças entre as categorias obedecem uma variação casual ($\chi^2 = 0,5233$ e $p - valor$>0.05), ou seja, não há evidência para afirmar que existe uma associação entre os níveis de concordância e a disciplina de preferencia dos usuário considerando um nível de significância de 95% (p-valor maior que 0.05 indica que deve-se aceitar H_0 e que não há evidência para afirmar que existe associação).

Concordância para disciplinas preferenciais iguais: Com o objetivo de detalhar os percentuais de concordância entre os pares de avaliadores de uma mesma disciplina preferencial (áreas primárias iguais), foram calculados os valores individuais de cada categoria. Apenas três (de um total de oito) disciplinas (ou áreas de atuação) possuem pares de avaliadores de uma mesma disciplina preferencial e que realizaram avaliações simultâneas, são elas: Ciências e Tecnologia, Educação e Humanas. A disciplina que possui um melhor nível de concordância entre os pares é a Ciências e Tecnologia com 15,2% de concordância satisfatória e 8,4% de concordância excelente. Para essas variáveis, não foi possível utilizar o qui-quadrado pois nem todos os níveis de concordância possuíam um mínimo de 5 casos.

Ao comparar os dados da Fig. 3 com relação ao percentual total e às disciplinas iguais, é possível verificar percentuais parecidos, isto é, os pares de avaliadores discordam fortemente das avaliações fornecidas. Mesmo quando são analisadas as disciplinas iguais separadamente os valores não diferem do que foi apresentado, ou seja, quando são analisados os pares de avaliadores, com disciplinas preferenciais iguais, para cada uma das disciplinas, quais sejam: Humanas, Educação e todas as outras, não se têm resultados melhores do que os apresentados, com exceção das áreas de Humanas e Educação que possuem um par de avaliador cada, concordando fortemente.

Concordância para disciplinas preferenciais distintas: Também foram avaliados individualmente os percentuais de concordância para os pares de usuários avaliadores pertencentes a disciplinas preferenciais distintas. Como apresentado na Fig. 3, os percentuais de concordância estão concentrados basicamente em um ICC classificado como pobre. As exceções são as seguintes combinações de áreas de atuação: 1) Educação x Ciência e Tecnologia, 2) Negócios X Humanas, e 3)Negócios X Ciência e Tecnologia. Para essas variáveis, não foi possível utilizar o qui-quadrado pois nem todos os níveis de concordância possuíam um mínimo de 5 casos.

Como pode ser visto na figura 3, o conjunto de dados analisado apresenta uma grande discordância entre as avaliações dadas pelos pares de avaliadores usuários. A discordância é levemente maior quando os usuários avaliadores pertencem a disciplinas preferenciais distintas, porém esta diferença não é estatisticamente significativa para afirmar dependência do nível de concordância com a categoria de disciplina.

B. Experimento 1B - Concordância entre pares considerando a atuação profissional

Também foi avaliado o nível de concordância com relação às classes de atuação profissional dos pares de avaliadores usuários. Conforme dito anteriormente, algumas dessas classes são: professor não universitário, estudante, profissional universitário, bibliotecário, consultor ou administrador. A Fig. 4 apresenta os percentuais de concordância para avaliadores que possuem a mesma atuação profissional e atuações profissionais distintas.

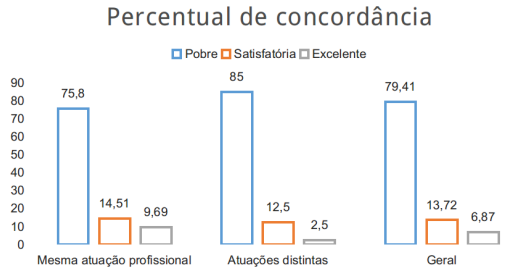

Fig. 4. Concordância geral considerando as áreas de atuação profissional

TABLE II
CONCORDÂNCIA CONSIDERANDO A MESMA ÁREA DE ATUAÇÃO PROFISSIONAL

Mesma atuação profissional	**Pobre**	**Satisfatória**	**Excelente**	**Total**
Professor não universitário	100% (8)			8
Estudante	71,69% (38)	16,98% (9)	11,33% (6)	53
Funcionário	100% (1)			1
Total de pares de avaliadores	75,80% (47)	14,51% (9)	9,69% (6)	62

Legenda: Valor em parênteses refere-se a quantidade de pares de avaliadores

A Fig. 4 mostra nível de concordância pobre entre os pares de avaliadores, independentemente de serem da mesma área de atuação profissional. Resultados muito parecidos com os resultados da seção anterior (ver Fig. 3).

Com o objetivo de aprofundar a análise, a Tabela II apresenta o percentual e a quantidade entre parênteses para as três classes de atuação profissional com pares de avaliadores que fizeram efetivamente alguma avaliação e são de uma mesma classe. Outras classes não formaram pares de avaliadores para um mesmo recurso. Como mostra a Tabela II, todos os pares de avaliadores usuários professores não universitários e funcionários apresentaram um nível de concordância pobre (ICC $<$ 0,4). Por outro lado, os avaliadores estudantes apresentaram um percentual de concordância maior, sendo que a soma dos percentuais de concordância satisfatório e excelente totaliza 28,31%, correspondendo ao grupo que mais concorda nas avaliações realizadas. Mesmo analisando os pares de avaliadores pela atuação profissional iguais e suas avaliações, ainda é possível observar que estes pares possuem divergência nas avaliações fornecidas.

A Tabela III mostra a concordância entre avaliadores de áreas de atuação profissional distintas. Como pode ser visto, 82,5% dos pares apresentam um nível de concordância pobre entre as avaliações dadas, com exceção de apenas 1 caso (Professor não universitário X Consultor), todos os demais casos com alguma concordância satisfatória ou excelente envolvem a classe Estudante. A Tabela III mostra uma situação similar a dos testes descritos anteriormente. O teste qui-quadrado apresentou um $p - valor$ = 0,77 (onde $p - valor$ > 0,05) indicando independência entre o nível de concordância e a

TABLE III
CONCORDÂNCIA CONSIDERANDO ÁREAS DE ATUAÇÃO PROFISSIONAL DISTINTAS

Atuação profissional		Pobre	Satisfatória	Excelente	Total
Prof. não universitário	Profissional universitário	100% (6)			6
Estudante	Funcionário	87,5% (7)	12,5% (1)		8
Estudante	Prof. não universitário	100% (3)			3
Estudante	Profissional universitário	77,78% (7)	22,22% (3)		10
Estudante	Administrador	75% (3)	25% (1)		4
Funcionário	Profissional universitário	100% (4)			4
Prof. não universitário	Funcionário	100% (1)			1
Profissional universitário	Bibliotecário	100% (1)			1
Prof. não universitário	Consultor	50% (1)		50% (1)	2
Bibliotecário	Prof. não universitário	100% (1)			1
Total de pares de avaliadores		85% (34)	12,5% (5)	2,5% (1)	40

Valor em parênteses refere-se a quantidade de pares de avaliadores

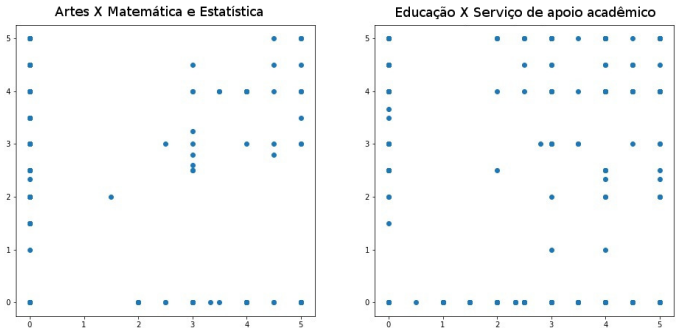

Fig. 5. Diagrama de dispersão da média das notas e as categorias

área de atuação profissional dos avaliadores usuários.

C. Experimento 2 - Correlação entre as médias das notas de cada categoria de disciplina

Neste experimento, foi verificado se existe correlação entre as médias das notas dadas para um recurso dependendo da categoria de disciplina preferencial dos usuários avaliadores. Na tabela resultante para a análise, cada linha representa um recurso, e cada coluna representa uma categoria de disciplina. Os valores das células são as médias das avaliações dadas para um recurso por usuários pertencentes a uma determinada categoria de disciplina (Educação, Humanas, Ciências Sociais, Ciências e Tecnologias, Serviços de apoio acadêmico, Negócios, Artes e Matemática e Estatística).

Neste contexto, foi realizada uma análise não paramétrica usando a correlação de *Spearman* (r_s) para analisar se existe associação entre as médias das avaliações de cada categoria de disciplina. A força de associação varia de -1 a 1 , sendo que valores próximos de zero indicam, não existe correlação.

As categorias de Artes x Matemática e Estatística, e Artes x Humanas foram as categorias que apresentaram a maior correlação entre as avaliações dadas pelos avaliadores, respectivamente, r_s = 0,15 ($p-valor$ = 4,76) e r_s = 0,09 ($p-valor$ = 2,35). Entretanto,P-valor foi maior que 0,05 indicando que não é possível afirmar uma associação entre as notas.

A Fig. 5 mostra um diagrama de dispersão para cada relação de categorias. Verificando a média das notas das categorias Artes e Humanas, existem diversos pontos que formam uma reta na vertical, indicando que as variáveis são independentes. A Fig. mostra uma leve dispersão dos pontos, mas a correlação observada é considerada fraca ($r_s = 0,09$). O $p - valor$ = $2,35$ $(> 0,05)$ não permite afirmar que as associações são significativas estatisticamente.

Como pode ser observado na Fig. 5, as médias de avaliações das categorias Artes x Matemática e Estatística não apresentam correlação, uma vez que o diagrama não mostra nenhum sinal de aglomeração de pontos em linha diagonal. O coeficiente de correlação para essas categorias é de r_s = 0,15 com $p-valor$ = 4,76, indicando uma correlação não significativa.

Análise similar se repete para todas as demais relações entre as categorias existentes. Na Fig. 5 o coeficiente obtido na correlação da disciplina de Educação x Serviços de Apoio Acadêmico é de r_s = 0,08 com P = 0,0001.

V. DISCUSSÃO

Após as análise dos resultados é possível responder às perguntas de pesquisa inicialmente propostas.

P1: Qual a concordância geral entre pares de avaliadores usuários de recursos educacionais digitais? **R1:** Os resultados mostram que os pares de avaliadores usuários possuem forte discordância com relação a qualidade dos recursos educacionais digitais que avaliaram simultaneamente. Dos 102 pares de avaliadores analisados, 79,41% possuem um nível de concordância pobre, 13,72% possuem um nível de concordância satisfatória e apenas 6,87% possuem um nível de concordância excelente.

P2: Qual o nível de concordância entre pares de avaliadores usuários considerando suas preferências por determinadas disciplinas e suas áreas de atuação profissional? **R2:** As estatísticas descritivas dos ICCs calculados apontam que existe um leve aumento da concordância entre os pares de avaliadores quando ambos pertencem a uma mesma disciplina preferencial. Sobre o nível de concordância considerando a área de atuação profissional dos avaliadores, o grupo de avaliadores que mais apresenta concordância em suas avaliações é o de estudantes (com aproximadamente 28% dos pares com concordância satisfatória ou excelente). De maneira geral, o único grupo de avaliadores que concorda de maneira satisfatória ou excelente com outros grupos também é o grupo de avaliadores de estudantes (exceção para o caso específico dos pares formados pelos grupos professor não universitário e consultor). Entretanto, é importante ressaltar que não foi possível encontrar uma associação significativa entre os diferentes grupos de avaliadores e o seu nível de concordância.

P3: É possível identificar alguma associação significativa entre as notas dadas pelos avaliadores usuários e as disciplinas de preferência dos avaliadores e dos recursos? **R3:** Os dados analisados apresentam uma leve correlação entre as notas e as disciplinas dos avaliadores e dos recursos, mas as correlações identificadas são muito fracas e sem significância estatística na maioria das vezes. As disciplinas que apresentaram as melhores correlações foram Artes x Humanas e Artes x Matemática e Estatística. A baixa associação entre as médias das pontuações dos recursos considerando as diferentes categorias de disciplinas é um indicativo de que esses diferentes grupos de avaliadores estão comunicando impressões diferentes com relação à qualidade dos recursos. Essa análise é complementar à análise anterior utilizando o ICC e auxilia na demonstração

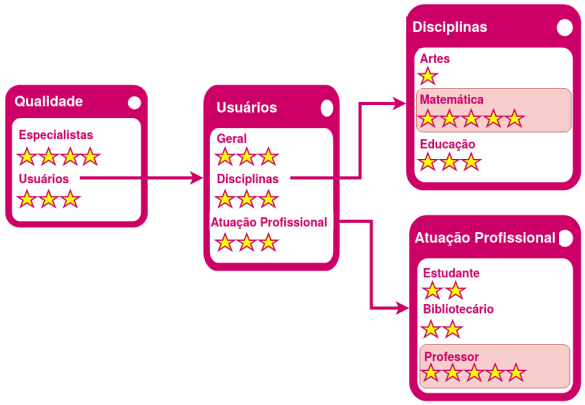

Fig. 6. Proposta de interface

de que grupos de usuários pertencentes a categorias de disciplinas diferentes tendem a pontuar os mesmos recursos de maneira distinta.

Os resultados obtidos mostram uma grande discordância entre as notas dadas pelos pares de avaliadores, e também pelos grupos de avaliadores pertencentes a diferentes categorias de disciplinas. Considerando que não existe um critério claro de avaliação para a comunidade de usuários no repositório estudado, não é possível precisar quais critérios estão sendo considerados nas avaliações realizadas, e é ainda mais difícil dizer qual a razão para os usuários fornecerem avaliações tão discordantes para um mesmo recurso. É importante ressaltar que a qualidade da revisão por pares já é criticada por diversos autores pela existência de grandes divergências sempre encontradas no processo de avaliação de trabalhos científicos [22], [23]. Neste contexto, esta pesquisa corrobora os trabalhos de [24], [25] pelo fato de conseguir detectar esta discordância entre as avaliações de pares de usuários, porém no contexto dos ROA.

Com o intuito de trazer soluções para as dificuldades apresentadas pelos resultados encontrados até aqui, busca-se apresentar melhorias na interface de apresentação da qualidade dos recursos dentro dos ROA, de maneira a contemplar as discordâncias existentes dentro do processo avaliativo. Desta forma, a grande preocupação é a forma com que as informações e funcionalidades são disponibilizadas e organizadas na interface.

Já existem estudos que buscam apontar as melhores estratégias para transmitir as informações na interface dos ROA de maneira simples e completa, e aperfeiçoando os sistemas de busca e navegação nestas plataformas, uma vez que o mal planejamento da interface do usuário pode levar ao impedimento de uso efetivo dos serviços fornecidos [26].

A Fig. 6 mostra uma nova forma de apresentação das informações sobre a qualidade dos recursos educacionais. O ROA deve, também, incluir as informações segmentadas pelas diferentes categorias existentes para os avaliadores (e.g. área de preferência - disciplina, e atuação profissional, entre outras existentes). Com isso, as diferentes opiniões sobre a qualidade dos OA são apresentadas de forma mais transparente/detalhada para os usuários da plataforma, permitindo que os mesmos tenham acesso direto às opiniões de qualidade mais próximas de seu interesse ou perfil. Os repositórios poderiam, também, personalizar a apresentação da qualidade de um recurso para um determinado usuário, considerando especificamente as suas características em comparação com as características dos diferentes tipos de avaliadores e as pontuações que os mesmos deram aos recursos (ressaltando na interface as pontuações mais similares com o seu perfil). No exemplo da Fig. 6, o usuário que está logado na plataforma e realizando a consulta de um determinado OA é da área de Matemática e atua profissionalmente como Professor (categorias grifadas em rosa nas janelas).

VI. Conclusões

A presente pesquisa utilizou dados recolhidos do repositório Merlot para medir a concordância e a correlação entre pares de avaliadores usuários que realizaram avaliações simultâneas sobre um mesmo conjunto de recursos educacionais digitais. Para a medição da concordância entre pares de avaliadores foi utilizada a medida ICC (Intra-class correlation) e, posteriormente foram analisadas as estatísticas descritivas para cada uma das classes de concordância existentes (pobre, satisfatória e excelente). Ainda, foram calculadas a concordância geral, e as concordâncias considerando as disciplinas de atuação preferencial e a atuação profissional dos avaliadores. Por fim, foram verificadas as correlações entre as médias das notas de cada categoria de disciplina.

Os resultados mostram uma discordância muito grande nas avaliações fornecidas pelos pares de avaliadores usuários, mesmo para aqueles pares que apontam as mesmas disciplinas ou áreas de atuação profissionais iguais. O percentual de correlação considerado pobre entre os pares de avaliadores apresentou-se maior do que 75% em todos os cenários estudados.

Uma das limitações do presente trabalho é a pequena quantidade de pares de avaliadores analisada e a consequente esparsidade das combinações de pares de avaliadores. Em contrapartida, cabe ressaltar que o Merlot é um dos repositórios mais disseminados atualmente e que conta com uma das maiores quantidades de recursos, usuários e avaliações fornecidas pela sua comunidade. Nesse sentido, ainda que a quantidade de pares possa ser considerada pequena, ela é representativa de um dos maiores e mais importantes ROA atualmente existentes. Outra limitação é a possibilidade da discordância ser resultante da falta de critérios objetivos de qualidade a serem seguidos pela comunidade de usuários. Considerando isso, pretende-se conduzir experimentos futuros com uma amostra de usuários que avaliem os recursos utilizando as mesmas dimensões de qualidade utilizadas pelo Merlot para a sua avaliação por especialistas, sendo elas: 1) facilidade de uso, 2) qualidade de conteúdo, e 3) potencial como ferramenta de ensino.

Gaona-García et al. [26] apontam que uma interface do usuário mal planejada pode impedir que os usuários utilizem os serviços oferecidos por determinado recurso (no nosso caso, um ROA). Sabe-se da importância de apresentar as informações referentes à avaliação dos OA de maneira mais completa, permitindo ao usuário entender o tipo de avaliador

que fornece as notas ao recurso e qual a sua área de atuação profissional. Assim, o usuário pode compreender a relação dos avaliadores com o material e aprofundar-se no entendimento da pontuação recebida. Neste contexto, foi apresentada uma proposta de interface com o objetivo de fornecer informações mais completas ao usuário e que contemple a discordância nas avaliações dos OA dentro dos ROA. Trabalhos futuros envolvem a implementação desse tipo de proposta em ROAs e a avaliação de sua aceitação por parte da comunidade de usuário, além de trabalhos complementares voltados à predição da avaliação do usuário [20] e da qualidade dos recursos [27].

AGRADECIMENTOS

Cristian Cechinel foi parcialmente financiado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico)[DT-2 - Bolsa de Produtividade em Desenvolvimento Tecnológico e Extensão Inovadora, proc.305731/2021-1].

REFERENCES

[1] C. Cechinel, S. Sánchez-Alonso, and E. García-Barriocanal, "Statistical profiles of highly-rated learning objects," *Computers & Education*, vol. 57, no. 1, pp. 1255–1269, 2011. [Online]. Available: https://doi.org/10.1016/j.compedu.2011.01.012

[2] K. Clements, J. Pawlowski, and N. Manouselis, "Open educational resources repositories literature review - Towards a comprehensive quality approaches framework," *Computers in Human Behavior*, vol. 51, pp. 1098–1106, 2015. [Online]. Available: https://doi.org/10.1016/j.chb.2015.03.026

[3] S. Downes, "Models for Sustainable Open Educational Resources," *Interdisciplinary Journal of Knowledge and Learning*, vol. 3, no. 1, pp. 29–44, 2007. [Online]. Available: https://doi.org/10.28945/384

[4] J. Atenas and L. Havemann, "Questions of quality in repositories of open educational resources: a literature review," *Research in Learning Technology*, vol. 22, Jul. 2014. [Online]. Available: https://journal.alt.ac.uk/index.php/rlt/article/view/1419

[5] M. S. e Cristian Cechinel e Vinicius Ramos, "Análise quantitativa da concordância entre pares de avaliadores de recursos digitais educacionais: um estudo de caso com dados do repositório merlot," *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, vol. 30, no. 1, p. 389, 2019. [Online]. Available: https://www.br-ie.org/pub/index.php/sbie/article/view/8743

[6] C. Cechinel and S. Sánchez-Alonso, "Analyzing associations between the different ratings dimensions of the merlot repository," *Interdisciplinary Journal of E-Learning and Learning Objects*, vol. 7, no. 1, pp. 1–9, 2011.

[7] J. Hylén, "Open educational resources: Opportunities and challenges," in *Proceedings of Open Education 2006: : Community, Culture, and Content*, Logan, UT, USA, 2006, acessado em 20 de abril de 2020: http://www.oecd.org/education/ceri/37351085.pdf.

[8] Shweta, R. Bajpai, and H. K. Chaturvedi, "Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods," *Journal of the Indian Academy of Applied Psychology*, vol. 41, no. 3, pp. 20–27, 2015.

[9] D. A. S. Matos, "Confiabilidade e concordância entre juízes: aplicações na área educacional," *Estudos em Avaliação Educacional*, vol. 25, no. 59, pp. 298–324, 2014.

[10] H. C. de Vet, C. B. Terwee, D. L. Knol, and L. M. Bouter, "When to use agreement versus reliability measures," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1033 – 1039, 2006. [Online]. Available: http://www.sciencedirect.com/science/article\/pii/S0895435606000291

[11] H. E. A. Tinsley and D. J. Weiss, "Interrater reliability and agreement." in *Handbook of applied multivariate statistics and mathematical modeling.* San Diego, CA, US: Academic Press, 2000, pp. 95–124.

[12] H. Vet, *Observer Reliability and Agreement*. American Cancer Society, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/\9781118445112.stat04910

[13] H. C. Kraemer, "Extension of the kappa coefficient," *Biometrics*, vol. 36, no. 2, pp. 207–216, 1980. [Online]. Available: http://www.jstor.org/stable/2529972

[14] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016. [Online]. Available: https://doi.org/10.1016/j.jcm.2016.02.012

[15] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, 3rd ed. John Wiley & Sons, Inc, 2003.

[16] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients." *Psychological methods*, vol. 1, no. 1, pp. 30–46, 1996.

[17] P. SHEA, S. MCCALL, and A. OZDOGRU, "Adoption of the multimedia educational resource for learning and online teaching (merlot) among higher education faculty: Evidence from the state university of new york learning network," in *MERLOT Journal of Online Learning and Teaching*, vol. 2, no. 3, 2006.

[18] C. Cechinel, "Objetos de Aprendizagem: introdução e fundamentos," *Editora UFABC*, vol. 1, pp. 71–77, 2015.

[19] C. Cechinel and S. Sánchez-Alonso, "Analyzing associations between the different ratings dimensions of the merlot repository," *Interdisciplinary Journal of E-Learning and Learning Objects*, vol. 7, no. 1, pp. 1–9, 2011.

[20] C. Cechinel, M.-Á. Sicilia, S. SáNchez-Alonso, and E. GarcíA-Barriocanal, "Evaluating collaborative filtering recommendations inside large learning object repositories," *Information Processing & Management*, vol. 49, no. 1, pp. 34–50, 2013.

[21] P. A. Barbetta, *Estatística aplicada às ciências sociais*, 2nd ed., U. F. d. S. Catarina, Ed., Florianópolis, 1994.

[22] L. Grayson and Q. Mary, "Evidence based policy and the quality of evidence: Rethinking peer review," *ESRC UK Centre for Evidence Based Policy and Practice Lesley*, 2002, Working Paper 7. [Online]. Available: http://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/assets/wp7.pdf

[23] J. Kelly, T. Sadeghieh, and K. Adeli, "Peer review in scientific publications : benefits , critiques , & a survival guide," *The Journal of the International Federation of Clinical CHemistry and Laboratory Medicine - JIFCC*, vol. 25, no. 3, pp. 227–243, 2014.

[24] S. Jenal, D. W. Vituri, G. M. Ezaías, A. Silva, M. Helena, and L. Caliri, "The peer review process : an integrative review of the literature *," *Acta Paulista de Enfermagem*, vol. 25, no. 5, pp. 802–808, 2012.

[25] M. Szklo, "Quality of scientific articles," *Rev. Saúde Pública*, vol. 40, pp. 30–35, 2006.

[26] P. A. Gaona-García, D. Martín-Moncunill, E. E. Gaona-García, A. Gómez-Acosta, and C. Monenegro-Marin, "Usability of big data resources in visual search interfaces of repositories based on kos," in *Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing*, ser. ICCBDC'18. New York, NY, USA: Association for Computing Machinery, 2018, p. 33–37. [Online]. Available: https://doi.org/10.1145/3264560.3264567

[27] C. Cechinel, S. Sánchez-Alonso, and E. García-Barriocanal, "Statistical profiles of highly-rated learning objects," *Computers & Education*, vol. 57, no. 1, pp. 1255–1269, 2011.