Universidade de Aveiro
Departamento de Electrónica, Telecomunicações e Informática
MEI

Análise da replicabilidade e reprodutibilidade de experiências de RI orientadas para o sistema

João Ferreira - 80041

20 de Janeiro de 2021

Trabalho realizado no âmbito da disciplina

Recuperação de Informação

1 Índice

2 Resumo

O tema escolhido consiste na análise da replicabilidade e reprodutibilidade em experiências de Recuperação de Informação (RI) orientadas para o sistema.

A replicabilidade e a reprodutibilidade dos resultados experimentais são preocupações primordiais em RI. Surge então um problema relacionado com o método em questão: não existe uma abordagem definida para avaliar o conteúdo reproduzido.

Esta monografia evidencia como estas duas métricas poderão ser avaliadas, através do coeficiente de Kendall τ, do algoritmo *Rank-Biased Overlap* (RBO) e da fórmula *Effect Ratio*, e a definição de reprodutibilidade como um processo nas ciências computacionais. Para além disso, é abordado um estudo que analisa a tendência existente em algoritmos de recomendação musical através da reprodutibilidade de um estudo já existente.

Após o término da análise de três artigos sobre esta temática ([1],[2] e [3]) foi possível concluir que a replicabilidade poderia ser avaliada através do coeficiente de Kendall τ e respetiva simplificação (que irá ser demonstrada posteriormente neste documento) e pelo algoritmo RBO. Concluiu-se também que a reprodutibilidade poderia ser analisada através da forma *Effect Ratio*. Para além destes dois aspetos conclusivos, foi possível afirmar que a reprodutibilidade é um processo, e não um objetivo, e que, relativamente às tendências que existem nos algoritmos de recomendação, os utilizadores têm apenas interesse limitado em itens populares e que os utilizadores que se interessam por itens não populares recebem recomendações piores que aqueles que se interessam por itens populares.

3 Introdução

Para a escrita desta monografia foram analisados três artigos científicos distintos, ambos sobre replicabilidade e reprodutibilidade em experiências de RI.

O primeiro dos artigos aborda como pode ser medida a replicabilidade e reprodutibilidade. Estas duas métricas constituem preocupações primordiais em RI e, no presente, não existe forma de as avaliar. Assim, surge um problema que apelidamos de *reproducibility crisis*, que se manifesta quando os investigadores não conseguem reproduzir e confirmar descobertas experimentais. Este problema envolve áreas como *Machine Learning* (ML) e Inteligência Artificial. Uma vez que estas duas áreas estão em constante expansão e integram um futuro relacionado com as tecnologias que se encontra cada vez mais próximo, surgiu a necessidade de correção deste problema. Para reproduzirem uma experiência, os investigadores tentam perceber como foi conduzida uma experiência e, após várias iterações, reproduzem-na. Infelizmente, não existem meios para medir esta reprodução.

O artigo seguinte afirma que a reprodutibilidade é um processo e não algo que possa ser descrito como uma conquista numa experiência. Existem diversas razões para considerarmos a reprodutibilidade como um processo recompensante, desde a demonstração de boa administração de recursos públicos (que financiam uma grande parte da pesquisa científica mundialmente) até ao fato da reprodutibilidade ser intrínseca ao próprio processo científico, aumentando a veracidade dos seus resultados e, consequentemente, das suas descobertas.

No último artigo selecionado, a pesquisa mostrou que os sistemas de recomendação são tipicamente tendenciosos para itens populares, o que torna os itens menos populares sub-

representados nas recomendações. Neste artigo são reproduzidas as análises de Abdollahpouri et al. [4] no contexto da recomendação musical. Foram investigados três grupos de utilizadores da plataforma Last.FM.

4 Análise de replicabilidade e reprodutibilidade

Como referido anteriormente, deparamo-nos com um problema que definimos como *Reproducibility Crisis*, manifestando-se quando os investigadores não conseguem reproduzir e confirmar descobertas experimentais.

Outro dos problemas existentes é que não está definida nenhuma coleção experimental focada especificamente na reprodutibilidade, algo que nos impede de desenvolver e comparar medidas para avaliar a extensão da reprodutibilidade alcançada.

É possível enfrentar estas adversidades. Inicialmente, podemos considerar diferentes medidas que permitem comparar os resultados experimentais em diferentes níveis:

i) Listas classificadas dos documentos recuperados;

ii) *Scores* das medidas de eficácia;

iii) Efeitos notados e diferenças significativas;

De seguida, deve ser produzido um *dataset* orientado para a reprodutibilidade, sendo utilizado para comparar estas medidas.

Resumidamente, a replicabilidade pode ser definida como algo que utiliza a mesma configuração experimental, sendo supervisionada e produzida por uma equipa diferente, enquanto que a reprodutibilidade difere da experiência original em ambos.

Para avaliarmos a replicabilidade começamos por calcular e ordenar os documentos pelo algoritmo de Kendall, que resulta na constante τ, pela seguinte fórmula:

$$\frac{P-Q}{\sqrt{(P+Q+U)(P+Q+V)}}$$

A fórmula descrita na imagem anterior é aplicada a dois conjuntos de documentos ordenados (*rankings)* aplicacados a uma colecção C, R e R', sendo que R' é o conjunto R replicado. Nesta fórmula cada uma das letras corresponde a um valor:

P – total de pares concordantes, isto é, ordenados pela mesma ordem nos dois conjuntos;

Q – total de pares discordantes, isto é, ordenados em posições opostas;

U – total de *ties* em R;

V – total de *ties* em R'.

Contudo, esta fórmula não é aplicável para dois conjuntos que não contenham o mesmo conjunto de documentos. A ideia principal é comparar a ordem de documentos no conjunto original R e no conjunto replicado R'.

Esta fórmula poderia ser simplificada para:

$$\frac{C-D}{C+D}$$

Nesta fórmula, cada letra possui um valor associado:

C – corresponde ao número de rankings com valor superior ao rank actual;

D – corresponde ao número de rankings com valor inferior ao rank actual:

Por exemplo, consideremos que R=[1,2,3] e R'=[1,2,4] e que pretendemos calcular a constante. Inicialmente, produzimos uma tabela com quatro colunas: R, R', C e D. Iniciamos o processo na primeira posição de R', ou seja na posição com o valor 1. Nas posições inferiores a esta, encontramos dois valores (2 e 4). Estes dois valores são ambos maiores que 1, ou seja, o valor de C seria igual a 2 e o valor de D igual a 0. Prosseguindo para a posição seguinte, 2, podemos concluir que o valor seguinte (4) é superior, pelo que C seria igual a 1 e D igual a 0. Este processo repete-se até ao fim da tabela, sendo que na última linha C e D são ambos iguais a 0.

Para concluir o cálculo pretendido, é necessário calcular o somatório da coluna com os valores de C e da coluna com os valores de D. Assim, o somatório da coluna com os valores de C seria igual a 3 e o somatório da coluna com os valores de D seria igual a 0. Finalmente, poderíamos afirmar que o valor resultante seria igual a 1 (3-0 a dividir por 3+0).

R	R'	C	D
1	1	2	0
2	2	1	0
3	4	0	0

Poderíamos considerar outro exemplo mais extenso, demonstrado pela tabela abaixo.

R	R'	C	D
1	12	0	11
2	2	9	1
3	3	8	1

4	4	7	1
5	5	6	1
6	6	5	1
7	7	4	1
8	8	3	1
9	9	2	1
10	10	1	1
11	11	0	1
12	1	0	0

Nesta tabela, o somatório de C seria igual a 45 e o somatório de D seria igual a 21. Concluindo, o valor resultante seria 0,37, aproximadamente.

Esta abordagem não nos informa das posições de documentos diferentes. Assim, para lidar com este problema é proposto utilizar o *Rank-Biased Overlap* (RBO), que assume que R e R' são conjuntos infinitos.

Para o cálculo do RBO poderíamos utilizar a seguinte fórmula:

$$\mathrm{RBO} = (\ 1 - \phi) \sum_{n=1}^{\infty} \phi^{i-1}.A_i$$

em que Φ pertence ao intervalo [0,1] e é um parâmetro para ajustar a medida e Ai é a proporção até à posição i, sendo igual à cardinalidade da intersecção dos dois conjuntos até i dividida por i.

Diferente da replicabilidade, para a reprodutibilidade o R e o R' não são obtidos da mesma coleção, ou seja, R não pode ser comparado com R'.

Assim , o coeficiente de Kendall e o RBO não podem ser aplicados. Focamo-nos então no *Effect Ratio*, colocando-se a seguinte questão: dados dois conjuntos, A e B, onde se sabe que A é mais abrangente que B numa determinada colecção C, poderá outro grupo reproduzir com melhorias para uma coleção diferente D?

Esta questão poderia ser descrita pela seguinte fórmula:

$$\mathrm{ER}\left(\Delta' M^C, \Delta M^C\right) = \frac{\overline{\Delta' M^C}}{\overline{\Delta M^C}} = \frac{\frac{1}{n_C} \sum_{j=1}^{n_C} \Delta' M_j^C}{\frac{1}{n_C} \sum_{j=1}^{n_C} \Delta M_j^C}$$

Na formula acima, nC corresponde ao número de tópicos em C e M_j^C ao vetor de scores no tópico j para a colecção C, sendo que Δ' corresponde à variação dos scores em A e B.

5 Reprodutibilidade como um processo

Habitualmente, em física, quando algo relacionado com fenómenos físicos é descoberto é pouco provável que se verifiquem alterações nesse fenómeno. Experiências subsequentes levam a novas questões e não podem ser consideradas como reprodutibilidade, *per se.*

A reprodutibilidade nas ciências computacionais tem características diferentes. Para este tópico, foi analisado um artigo, escrito em 2019, no qual foi replicado um desafio *Open Source* de 2015 e no qual é colocada a questão: quatro anos depois estes artefactos computacionais ainda serão funcionais? Nos resultados exibidos neste artigo é possível concluir que os autores não foram capazes de replicar a maioria das execuções num ambiente computacional moderno.

Este teste foi realizado por dois motivos:

- As descobertas nas ciências computacionais não são sempre expressas em algum contexto computacional como, por exemplo, o algoritmo A é mais rápido que o algoritmo B. Surge assim uma questão: com estruturas de dados residentes em memória, a descoberta ainda se mantém?
- Preservação de software a longo prazo, especialmente para artefactos que possuem valores históricos.

6 Tendência em sistemas de recomendação

O último artigo analisado aborda a tendência de popularidade na recomendação musical de um estudo de reprodutibilidade. A pesquisa mostrou que os sistemas de recomendação são, geralmente, tendenciosos para conteúdo popular, o que torna o conteúdo menos popular sub-representado nas recomendações.

Neste artigo foi reproduzida uma análise de recomendação musical, sendo que foram investigados três grupos de utilizados da plataforma Last.FM. Estes grupos foram categorizados da seguinte forma:

i. Utilizadores pouco convencionais;

ii. Utilizadores de médio porte;

iii. Utilizadores de grande porte.

De acordo com o artigo, os investigadores descobriram que algoritmos de recomendação de última geração favorecem itens populares também no domínio da música. No entanto, a métrica de popularidade do grupo de médio porte produz resultados diferentes para músicas e filmes, provavelmente devido ao maior número de artistas musicais no *dataset* utilizado.

Os autores Dominik Kowald, Markes Sheld e Elisabeth Lex mostraram que os algoritmos de recomendação atuais têm tendência para mal servir os utilizadores que gostam de itens não populares, ao reproduzirem um estudo no domínio da música.

Colocaram duas questões:

1) Até que ponto os utilizadores ou grupos de utilizadores estão interessados em músicas populares?
2) Até que ponto a métrica de popularidade dos algoritmos de recomendação afeta os utilizadores com inclinações para a música *mainstream*?

Para este estudo foram considerados 3000 utilizadores, 1000 para cada grupo. Os autores concluíram que um número muito reduzido de artistas são ouvidos por muitos utilizadores enquanto que a maioria dos artistas são ouvidos por poucos utilizadores.

De seguida investigaram se há correlação entre o perfil dos utilizadores (número de artistas) e a popularidade dos artistas no perfil do utilizador e encontram uma correlação igual a 0.965, como seria de esperar, uma vez que a probabilidade de ter artistas populares no perfil aumenta com o número de itens no perfil.

No entanto, correlacionando a popularidade média dos artistas no perfil do utilizador sobre o tamanho do perfil encontraram uma correlação negativa (-0.372), o que significa que os utilizadores com um tamanho de perfil menor tendem a ouvir mais artistas populares.

Respondendo à primeira questão, descobriram que um terço dos utilizadores tem pelo menos 20% de artistas não populares nos seus perfis e, portanto, também estão interessados em música popular.

Para além dos dados já referidos nos parágrafos anteriores, os autores estudaram também a métrica de popularidade em algoritmos de recomendação de música de última geração. Utilizando o Surprise4, *toolkit* de Python, formularam as recomendações musicais como um problema de previsão de classificação, onde previram a preferência de um utilizador por um artista. Consideraram cinco algoritmos:

1. Random;
2. MostPopular;
3. User Item Average;
4. User KNN;
5. User KNN Average;

Representaram a correlação de popularidade do artista e a frequência com que os cinco algoritmos recomendam esses artistas. Para todos os algoritmos, exceto o Random, encontraram uma correlação positiva, o que significa que itens populares são recomendados com mais frequência do que itens não populares. Como esperado, este efeito é mais evidente para o algoritmo MostPopular e não está presente no algoritmo Random.

Os autores concluíram que:

1. Os utilizadores têm apenas um interesse limitado em itens populares;

2. Os utilizadores interessados em itens não populares recebem recomendações piores do que os utilizadores interessados em itens populares;

7 Conclusões

Concluindo, existem diversas formas para avaliarmos a reprodutibilidade e replicabilidade de resultados experimentais, como o coeficiente de Kendall, o *Rank-Based Overlap* e o *Effect Ratio*. Estas abordagens ajudam-nos a ter uma melhor perceção do nível de reprodutibilidade.

Podemos concluir também que a reprodutibilidade pode e deve ser abordada como um processo e não um objetivo, de modo a alcançarmos a preservação de software a longo prazo, especialmente para artefactos que possuem valor histórico.

Finalmente, foi possível concluir que os algoritmos de recomendação atuais têm tendência para mal servir utilizadores que gostam de itens não populares e que apenas um número muito reduzido de artistas são ouvidos por muitos utilizados, enquanto que a maioria dos artistas são ouvidos por poucos utilizadores. Conclui-se também que os utilizadores interessados em itens não populares recebem recomendações piores do que os utilizadores interessados em itens populares.

8 Referências

[1] Breuer, T., Ferro, N., Fuhr, N., & Sakai, T. (2020). How to Measure the Reproducibility of System-oriented IR Experiments. *SIGIR2020*, 1–15. https://www.researchgate.net/publication/342956521_How_to_Measure_the_Reproducibility_of_System-oriented_IR_Experiments

[2] Lin, J., & Zhang, Q. (2020). Reproducibility is a Process, not an Achievement: The Replicability of IRReproducibility Experiments. *ECIR 2020*, 1–7. https://www.researchgate.net/publication/340573305_Reproducibility_is_a_Process_Not_an_Achievement_The_Replicability_of_IR_Reproducibility_Experiments

[3] Kowald, D., Schedl, M., & Lex, E. (2020). The Unfairness of Popularity Bias in MusicRecommendation: A Reproducibility Study. *ECIR 2020*, 1–7. https://arxiv.org/pdf/1912.04696.pdf