Ferramentas de Análise Estatística e Correlação para Sistemas de Recuperação da Informação

Khaíck Oliveira Brito[1] **, Rodrigo Tripodi Calumby**[1]

[1]Universidade Estadual de Feira de Santana (UEFS)
Feira de Santana - Bahia - Brasil

khaickbrito@gmail.com, rtcalumby@uefs.br

Abstract. *In order to compare information retrieval systems using efficacy measures, it is necessary to use a variety of tools and programs to perform efficacy, statistical analyzes and correlations. Thus, the comparison process ends up being long and complex due to the non integration of these tools. In order to overcome these difficulties, which directly affect researchers and developers, in this work were added analyzes, statistical tests with different libraries, heat map as a new possibility of visualizing results and tables of critical values to the range of resources of the tool AnalyzIR to solve multiple needs an integrated tool.*

Resumo. *Para se comparar sistemas de recuperação da informação utilizando medidas de eficácia faz-se necessário utilizar uma variedade de ferramentas para realizar computo de eficácia, análises estatísticas e correlações. Assim sendo, o processo de comparação acaba sendo longo e complexo devido à não integração dessas ferramentas. Visando superar essas dificuldades, as quais atingem diretamente pesquisadores e desenvolvedores, nesse trabalho foram acrescidos análises, testes estatísticos com diferentes bibliotecas, mapa de calor como uma nova possibilidade de visualização de resultados e tabelas de valores críticos à gama de recursos da ferramenta AnalyzIR visando solucionar múltiplas necessidades por meio de uma ferramenta integrada.*

1. Introdução

Notadamente, cresce a quantidade de informações que são inseridas no universo digital, principalmente na Internet. É possível observar o surgimento de novos aplicativos, sistemas, sensores e outras ferramentas que coletam e geram dados, esses que por vezes precisam ser armazenados e encontrados. Os responsáveis por tais resgates são os sistemas de *Recuperação da Informação* (RI). Tais sistemas precisam ter bons algoritmos para recuperar os dados relevantes de maneira eficiente e eficaz. Dessa forma, podem existir sistemas que utilizam de diferentes estratégias para recuperar essas informações, tal que pesquisadores e desenvolvedores buscam comparar esses sistemas para escolher qual utilizar, melhorá-los ou desenvolver novos sistemas.

No entanto, para esta atividade é necessário utilizar um conjunto de ferramentas para mensurar eficácia, realizar análise estatística, fazer análises de correlação para diferentes cenários e construir modelos visuais para apresentação dos resultados. Em especial, para garantir o rigor científico na análise comparativa de eficácia de diferentes sistemas,

é necessária a utilização de métodos para determinação da significância estatística considerando a execução dos algoritmos em múltiplas bases de dados e consultas. De modo geral, estas são tarefas trabalhosas dado que as ferramentas utilizadas frequentemente consomem e geram dados em diferentes formatos e possuem um conjunto limitado de funcionalidades. Neste trabalho, foi realizada a integração de funcionalidades de análise de significância estatística e correlação à AnalyzIR, ferramenta que tem sido desenvolvida na UEFS e que visa disponibilizar ao usuário um ambiente integrado de análise de eficácia de sistema de recuperação de informação.

2. Metodologia

Visando complementar a ferramenta, foram adicionados novas funcionalidades, sendo os testes de *Mann-Whitney U* [Carmo 2019b], Correlação de *Pearson* [Carmo 2019a], Correlação de *Spearman* [Carmo 2019a], Coeficiente de *Jaccard* [Rodrigues 2019] e também foram aprimorados testes já existentes, como *Wilcoxon's Signed Rank Test* [Carmo 2019c] e Teste t de *Student* [Alves 2019]. Os testes de significância estatística formulam duas hipóteses, nula e alternativa, que serão consideradas para concluir acerca das amostragens às quais o teste será aplicado, como por exemplo avaliar a existência de uma diferença estatisticamente significativa entre as médias dessas amostras.

Os testes de *Wilcoxon* e *Mann-Whitney* U podem ser aplicados para se comparar duas amostragens relacionadas ou emparelhadas visando verificar diferenças significativas, caso existam, entre os comportamentos dessas amostragens, podendo concluir se há chance de elas pertencerem à mesma população ou não. Ambos computam a diferença entre os valores emparelhados, a magnitude dessa diferença, em seguida ordenam essas diferenças e definem postos sinalizados. Por fim, aplica-se a fórmula estatística do teste em questão, encontrando um valor a ser comparado com o tabelado considerando o grau de liberdade e nível de significância escolhidos. Caso o valor encontrado seja superior ao tabelado, é indicado se rejeitar a hipótese nula de que as amostragens se assemelham ou suas médias são estatisticamente equivalentes.

O cálculo do valor t de *Student* se faz de maneira específica mediante a igualdade ou não da variância e da igualdade ou não entre os tamanhos das amostragens. Esse valor encontrado será comparado, posteriormente, aos valores presentes na tabela de valores críticos de t, considerando determinado grau de liberdade e valor de confiança.

O coeficiente de correlação de *Jaccard* é uma estatística usada para comparar conjuntos finitos e calcular um índice para a similaridade entre eles. O valor encontrado utilizando operações entre os conjuntos a serem comparados, fica inserido num intervalo $[0, 1]$, em que quanto maior o coeficiente mais alta a similaridade. O coeficiente de correlação de *Pearson* busca avaliar a existência de uma relação linear entre as variáveis apresentadas nas amostras. Através do cômputo de valores de variância e covariância, o teste de *Pearson* gera um valor p que indica o grau da correlação encontrado no intervalo $[-1, 1]$, em que valores próximos de 1 expressam uma correlação direta entre as variáveis analisadas, -1 para uma correlação inversa e 0 para uma não dependência entre elas. O teste de correlação de ordem de posto de *Spearman*, também como em *Pearson*, procura encontrar e apresentar o tipo de relação existente entre as variáveis contínuas ou ordinais apresentadas a partir de valores de covariância e desvio padrão.

Especificamente para efetuar o teste de *Spearman* devidamente para sistemas de

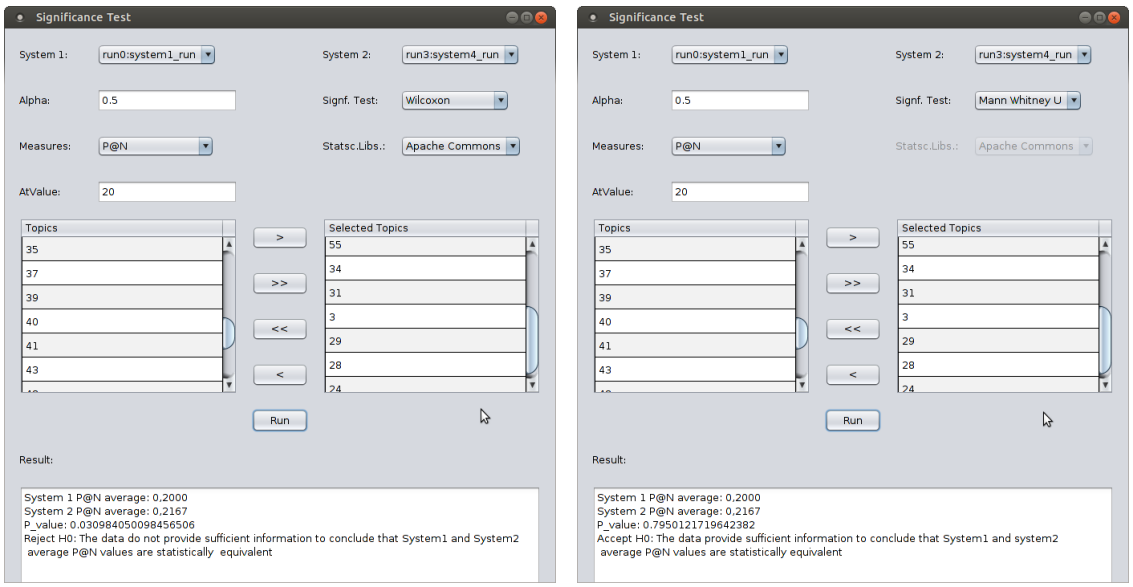
RI, foi necessário desenvolver e aplicar um método de adaptação para rankings que não compartilham de todos os elementos entre si, ou seja, que não tenha totalidade de intersecção nos elementos dos rankings, estimando o posto de um elemento não presente em um ranking, mas presente no outro, sendo mais distante o posto atribuído ao elemento ausente que se encontra mais próximo do topo do ranking ao qual ele pertence, de modo a permitir o cômputo do coeficiente. Detalhes do método desenvolvido são descritos na próxima seção.

Tratando-se da aplicação desses testes em relação à ferramenta e à análise de sistemas de RI, os parâmetros necessários para se efetuar devidamente os testes são:

- Testes de *Wilcoxon* e *Mann-Whitney U*: Escolha de dois sistemas a serem avaliados, um valor para o grau de confiança, e.g., 0.05 (95%), uma medida de avaliação, um valor de profundidade para os rankings e uma ou mais consultas.
- Teste t de *Student*: Escolha de dois sistemas a serem avaliados, um valor para o grau de confiança , e.g., 0.05 (95%), uma medida de avaliação, um valor de profundidade para os rankings e uma ou mais consultas.
- Teste de *Spearman* e coeficiente de *Jaccard*: Escolha de um ou mais sistemas, uma ou mais consultas e um valor de profundidade
- Teste de *Pearson*: Escolha de dois sistemas a serem avaliados, uma medida de avaliação, um valor de profundidade para os rankings e uma ou mais consultas.

3. Resultados e Discussões

A interface para o teste de *Wilcoxon* é exibida na Figura 1-a e o teste de *Mann Whitney U* na Figura 1-b. Elas apresentam os campos necessários para suas execuções, bem como suas regiões de resultados.



(a) Interface para o teste de *Wilcoxon*

(b) Interface para o teste de *Mann Whitney U*

Figura 1. Interfaces gráficas para execução de testes não paramétricos

Existe a possibilidade de três versões para o teste de *Wilcoxon*: Apache Commons Math (ACM) [Commons 2018], Java Statistical Classes (JSC) [Java 2017] e Mackenzie [MacKenzie 2018], e duas para Mann *Whitney U*, ACM e JSC, que podem ser escolhidas na interface apresentada. As interfaces para os testes t de Student e correlação de Pearson se assemelham à apresentada na Figura 1.

O método desenvolvido para adaptação de vetores, citado na seção anterior e ilustrado na Figura 2, o qual visa, por meio de sua adaptação, permitir a execução do teste suprindo necessidade de se possuir os mesmos elementos nos vetores a serem usados como base do teste, se dá da seguinte maneira:

1. Forma-se um conjunto com a união dos elementos dos rankings A e B, sendo A e B conjuntos com tamanhos iguais.
2. Criam-se dois novos vetores, A1 e B1, correspondentes à A e B. Para cada elemento dessa união:
3. Verifica-se se ele existe em ambos os vetores originais A e B
4. Caso exista em algum deles, o valor da sua posição no seu vetor de origem será guardado no vetor correspondente ao vetor original em que foi encontrado
5. Caso somente exista em um dos dois rankings originais, o posto atribuído a esse elemento no vetor correspondente ao ranking ao qual o elemento não faz parte será calculado pela equação $Posto = 2T - P$, onde T corresponde ao tamanho do ranking original e P é a posição ocupada pelo elemento no ranking. original.

A equação do quinto passo foi pensada para supor uma posição para um elemento não existente no ranking de forma ponderada. Quanto mais alto o elemento está no ranking original, mais baixo ele estará no ranking o qual ele está sendo comparado. A Figura 2 demonstra a aplicação do método de adaptação em questão onde A e B são vetores de tamanhos iguais porém com alguns elementos iguais (A, B e C) e outros diferentes (D e E).

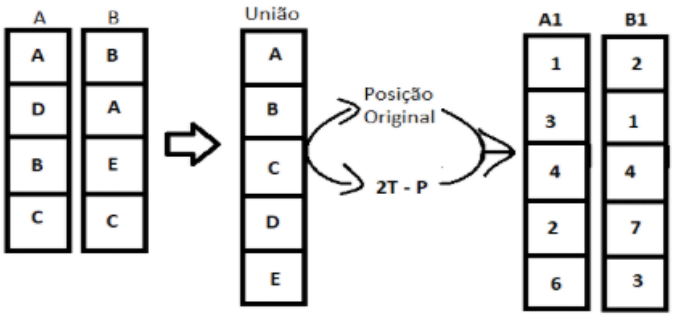
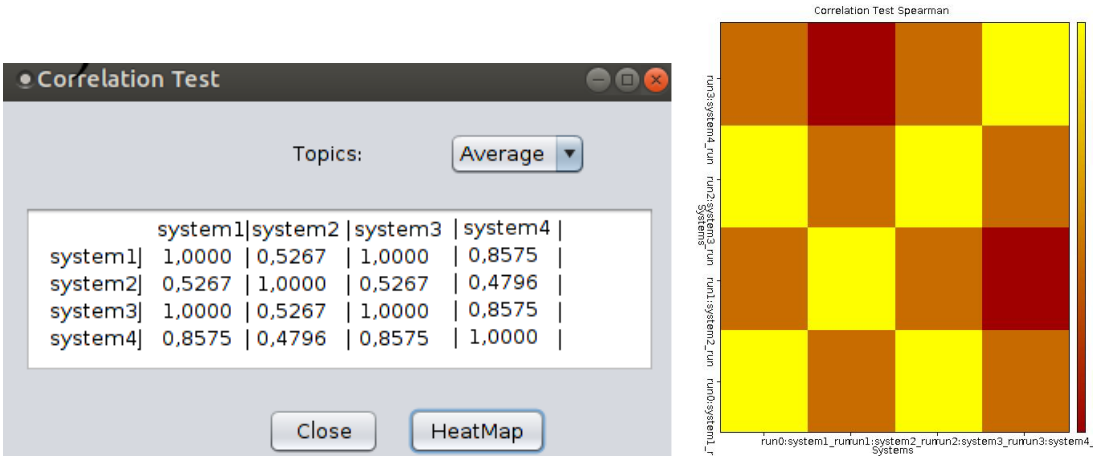


Figura 2. Exemplo de adaptação de *rankings*.

A visualização dos resultados dos testes de Spearman e Jaccard pode ser feita através de uma matriz de correlação, Figura 3-a, ou por meio de mapa de calor, Figura 3-b, o qual permite edições dos títulos, legendas e escala de cores. É possível realizar a exportação desse mapa de calor em formatos PNG e JPG.



(a) Matriz de correlação (b) Mapa de calor gerado e exportado

Figura 3. Exibição de resultados

Visando auxiliar nas conclusões estatísticas acerca dos resultados obtidos, são disponibilizadas na própria ferramenta tabelas de valores críticos t, Figura 4, e Spearman.

Student's t-Test

df (N-1)	0.90	0.95	0.975	0.99	0.995	0.999
1	3,078	6,314	12,706	31,821	63,657	318,313
2	1,886	2,920	4,303	6,965	9,925	22,327
3	1,638	2,353	3,182	4,541	5,841	10,215
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,782
8	1,397	1,860	2,306	2,896	3,355	4,499
9	1,383	1,833	2,262	2,821	3,250	4,296
10	1,372	1,812	2,228	2,764	3,169	4,143

Figura 4. Tabela t de valores críticos.

4. Conclusão

Com o desenvolvimento desse trabalho, foi possível contemplar os resultados obtidos satisfatoriamente, incluindo o aprimoramento da ferramenta AnalyzIR, a qual se encontra em fase final de desenvolvimento mas será disponibilizada em breve, por meio da incorporação do módulo estatístico desenvolvido nesse projeto, a qual permitirá aos pesquisadores e desenvolvedores que fizerem uso dos recursos da ferramenta, a qual será divulgada de forma aberta e gratuita, efetuarem testes com bibliotecas conceituadas e suporte a múltiplos sistemas, permitindo um maior rigor para conclusões científicas para

embasarem ainda mais seus trabalhos, visualizações de resultados de forma mais intuitiva através de mapas de calor com legendas e cores customizáveis junto à exportação de resultados e tipos diferentes de arquivos, para se adequar às diferentes necessidades.

Referências

Alves, M. C. (2017 (acessado Março 09, 2019)). *Test t de Student. http://cmq.esalq.usp.br/wiki/lib/exe/fetch.php?media=publico:syllabvs:lcf5759a:teste$_t$.pdf*.

Carmo, V. (2012 (acessado Março 09, 2019)a). *Correlação. http://www.inf.ufsc.br/ vera.carmo/Correlacao/Correlacao$_P$earson$_S$pearman$_K$endall.pdf*.

Carmo, V. (2012 (acessado Março 09, 2019)b). *Teste de Mann-Whitney. http://www.inf.ufsc.br/ vera.carmo/Testes$_d$e$_H$ipoteses/Testes$_n$ao$_p$arametricos$_M$ann−Whitney.pdf*.

Carmo, V. (2012 (acessado Março 09, 2019)c). *Teste de Wilcoxon. http://www.inf.ufsc.br/ vera.carmo/Testes$_d$e$_H$ipoteses/Teste$_n$ao$_p$arametrico$_W$ilcoxon.pdf*.

Commons, A. (2016 (acessado Agosto 15, 2018)). *Commons Math: The Apache Commons Mathematics Library. http://commons.apache.org/proper/commons-math/*.

Java (2017 (acessado Outubro 15, 2017)). *Java Statistical Class. http://www.jsc.nildram.co.uk*.

MacKenzie, S. (2016 (acessado Agosto 15, 2018)). *Class Wilcoxon Signed Rank by Scott MacKenzie. http://www.yorku.ca/mack/HCIbook/stats/WilcoxonSignedRank.html*.

Rodrigues, V. (2018 (acessado Março 09, 2019)). *Índice de similaridade de Jaccard e dendrograma no R. https://medium.com/bio-data-blog/%C3%ADndice-de-similaridade-de-jaccard-e-dendrograma-no-r-eba05d9d313a*.