Problemas de Classificação com Agregação de Ranking

Palavras-Chave: Rankings, Ranking Aggregation, Consensus Problems

Autores(as):

Pedro Carvalho Cintra, IC – UNICAMP

Prof(ª). Dr(ª). Flávio Keidi Miyazawa (orientador(a)), IC - UNICAMP

1 Introdução

Problemas de ranqueamento, particionamento e agregação de informações permeiam as diversas áreas do conhecimento de diversas formas e o interesse por agregar dados é o ponto central deste trabalho. Nesse sentido, o presente relatório diz respeito a um estudo aprofundado sobre a definição e estratégias de solução para esses tipos de problemas, com o foco na identificação de como podemos perceber um consenso entre diferentes opiniões através do particionamento e agregação dessas opiniões.

Nesse sentido, dado a grande parcela de problemas que envolvem a construção de consensos, é de interesse do saber científico entender como agregar informação e modelar computcionalmente esse tipo de operção entre um conjunto de dados. Em especial, estamos interessados em agregar opiniões de diferentes indivíduos em um âmbito geral. Assim sendo, dado um conjunto de itens $S = \{e_1, e_2 \ldots, e_n\}$, os quais podem ser classificados diversificadamente, e um conjunto de eleitores $V = \{v_1, v_2 \ldots, v_N\}$, onde cada eleitor deve demonstrar uma opinião sobre esses itens, o problema trata-se de encontrar um conjunto de opiniões que seja um consenso entre os eleitores.

Inicialmente, pensando da maneira mais genérica no problema de consenso entre opiniões, um eleitor poderia fazer uma comparação par a par sobre os itens dizendo por exemplo se um dado item e_i é melhor, pior ou simplesmente não sabe opinar do que um item e_j, sendo que "melhor" ou "pior" seja apenas um conceito opinativo dos eleitores baseado em alguma métrica que faça sentido, como, por exemplo, uma revista científica é melhor que outra, na opinião de um autor. Tendo em vista esse cenário, de imediato algumas conclusões podem ser inferidas se modelarmos o problema como um grafo direcionado onde um vértice representa um item e existe um aresta direcionada de e_i para e_j se o eleitor julga e_i melhor do que e_j. Por exemplo, um caminho nesse grafo, levando em conta apenas um eleitor, gera uma espécie de *ranking*, o que é algo bem mais específico e possui maior quantidade de informação do que relações "soltas" par a par pelo grafo.

Nesse sentido, o espaço de estudo se voltou para conhecer melhor o problema conhecido como *Ranking Aggregation Problem*, que já possui diversos e importantes resultados. Portanto, redefinimos o problema simplificando-o da seguinte forma: dado um conjunto de itens $S = \{e_1, e_2 \ldots, e_n\}$ e um conjunto de *rankings* $P = \{\pi_1, \pi_2 \ldots, \pi_N\}$, onde um *ranking* π_i é uma ordem sobre um subconjunto $D \subseteq S$, isto é, por exemplo, $\pi_1 = (e_2, e_1, e_{n-2}, e_{n-5}, e_{n-4})$ indicando que e_2 é melhor que e_1, que é melhor que e_{n-2}, assim sucessivamente, fazendo com que o e_{n-4} sejá o pior colocado em D segundo π_1. Desse modo, a redefinição do problema trata

de encontrar uma permutação σ sobre S que represente o consenso entre todos os *rankings*. Nessa lógica, existem diversos desdobramentos sobre esse problema, além de diferentes métricas para definir o que seria esse melhor consenso.

Tendo suas raízes no âmbito combinatório, esse problema foi amplamente discutido sob uma perspectiva matemática. Em particular, Kemeny [1, 2] propôs um critério de otimização para escolher o melhor consenso entre um conjunto de rankings. Essa métrica basicamente se trata de minimizar a quantidade de desacordos em pares entre um *ranking* escolhido como o consenso e todos os outros *rankings* dos eleitores. Assim sendo, dado o conjunto de *rankings* $P = \{\pi_1, \pi_2 \ldots, \pi_N\}$ o intuito é de minimizar a função objetivo $\sum_{i=1}^{N} d(\sigma, \pi_i)$, onde $d(\sigma, \pi)$ é quantidade de pares que estão que estão em desacordo, ou seja, ordenadas de maneira contrária, entre os dois *rankings*. Formalmente, definindo que $i \prec_\pi j$ significa que o item i precede o item j no *ranking* π, temos que

$$K(\sigma) = \sum_{i=1}^{N} d(\sigma, \pi_i) \tag{1}$$

onde

$$d(\sigma, \pi) = |\{(i, j) : i < j, (i \prec_\sigma j \wedge j \prec_\pi i) \vee (j \prec_\sigma i \wedge i \prec_\pi j)\}| \tag{2}$$

Essa distância leva o nome de *Kendall-tau distance* ou *bubble sort distance*, devido a sua correlação com o respectivo processo de ordenação. É notório, pela pesquisa realizada, que essa é uma das métricas de comparação mais usadas no que se trata de algoritmos relacionados à agregação de *rankings*. É conhecido que encontrar um *ranking* σ que minimize a função de Kemeny é um problema *NP-Hard* [3] e permanece *NP-Hard* mesmo quando existem apenas quatro *rankings* para serem agregados [4]. Sendo assim, já que uma de suas principais métricas certamente não possui uma solução polinomial, é natural que grande parte das abordagens para resolver o problema de agregação de *rankings* sejam em sua maioria heurísticas e/ou algoritmos de aproximação.

2 Metodologia

Tendo em vista esse contexto, o presente relatório apresenta a comparação entre diferentes abordagens através de sua implementação e teste sobre banco de dados contendo ranqueamentos reais, tendo como método de comparação a *Kendall-tau distance*.

Sendo assim, foram implementadas três diferentes heurísticas para o *Ranking Aggregation Problem*. Uma das abordagens - muito citada no meio - é a utilização de Cadeias de Markov, por Dwork et al.[4, 5]. Empregada no âmbito de *metasearch* e *spam* em motores de busca, essa técnica diz respeito de, essencialmente, modelar o *Rank Aggregation Problem* como um problema de estados em uma Cadeia de Markov, onde os estados da cadeia seriam o conjunto de candidatos que estão sendo ranqueados, a matriz de transição seria construída a partir dos *rankings* fornecidos e a distribuição estacionária forneceria a ordem consenso entre os itens. Quatro tipos de matrizes de transição foram estabelecidas pelos autores, todas com o mesmo enfoque mas com características diferentes, o que é ainda mais relevante a título de comparação entre os resultados.

Como segunda alternativa, o trabalho utilizado foi o fornecido por Aledo et al. [6], onde ele utiliza de uma generalização das matrizes de precedência, uma alternativa de pontuação que conta o percentual de vitórias par a par (equivalente ao *Borda Count*), para tratar de casos mais gerais da agregação de *rankings*, no caso, permitindo *ties* e informações parciais nos *rankings*. Ou seja, os *rankings* não necessariamente precisam ser permutações do conjunto de itens e ainda possuir itens "empatados" no sistema de preferência. Aledo et al. faz isso introduzindo um conceito chamado *Extension Sets*, que se trata, essencialmente, de um conjunto, construído a partir de cada ranqueamento, que possui todas as permutações do total de itens que concordam com as ordens estabelecidas no *ranking*. Utilizando essas informações na matriz de precedência é possível, assim, gerir as informações incompletas.

Por último, dos algoritmos implementados, a abordagem apresentada por Xiao et al. [7], constrói um grafo de competição entre itens com base nas comparações observadas nos rankings. As arestas direcionadas representam vitórias de um item sobre outro, ponderadas pela frequência relativa dessa preferência. O índice ROID (*ratio of out- and in-degree*) é então definido para cada item como a razão entre seu grau de saída e grau de entrada no grafo. O consenso final é obtido ordenando os itens de acordo com esses índices.

Além disso, foi implementada uma estratégia de melhoramento introduziada por Dwork et al.[4, 5] chamada de *Local Kemenization*. Essa técnica realiza trocas locais de pares adjacentes sempre que isso reduzir a distância de Kendall total, buscando refinar o consenso porém sem necessariamente perder as características da abordagem inicial.

Todos os algoritmos foram implementados em C++ e compilados com compilador gcc version 14.2.1 20240912 (Red Hat 14.2.1-3) (GCC) e, como conjunto de dados para análise, foram utilizados as seguintes dez coleções:

1. **F1-2013**: Posições de pilotos de corrida na Fórmula 1 de 2013.
2. **F1-2014**: Posições de pilotos de corrida na Fórmula 1 de 2014.
3. **Sushi**: Ranqueamento feito por participantes de um questionário sobre pratos de *Sushi*.
4. **Tour of France**: Posicões de ciclitas no *Tour of France* de 2013.
5. **ATP-Men-50**: Top 50 tenistas pela ATP (2014).
6. **ATP-Women-50**: Top 50 tenistas pela ATP (2014).
7. **ATP-Men-100**: Top 100 tenistas pela ATP (2014).
8. **ATP-Women-100**: Top 100 tenistas pela ATP (2014).
9. **ATP-Men-200**: Top 200 tenistas pela ATP (2014).
10. **ATP-Women-200**: Top 200 tenistas pela ATP (2014).

3 Resultados e Discussão

No que tange os resultados obtidos pelas implementações, é primeiramente exibida aplicação das diferentes Cadeias de Markov definidas por Dwork et al.[4, 5] como mostrado na Figura 1. Devido ao fato de que os resultados estão em escalas diferentes - valores de *Kendall-tau distance* muito discrepantes - dependendo do conjunto de dados, foi feita a normalização para cada conjunto de dados dividindo os valores dos algoritmos representados pelo resultado obtido pelo pior dos algoritmos (nesse primeiro caso MC1).

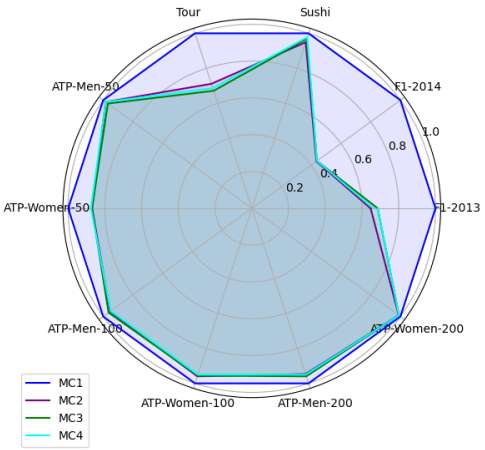

Figure 1: Comparação entre as 4 Cadeias de Markov

Analisando os restultados, percebe-se que MC2, MC3 e MC4 possuem desempenhos muito semelhantes e relativamente melhores quando comparados com MC1. Por isso usaremos MC4 como represente da abordagem de Cadeias de Markov para comparação com outras abordagens.

Assim sendo, é evidenciado pela Figura 2, que dentre todos os algoritmos propostos o que melhor performou dentre os escolhidos foi a aplicação do Borda Count utilizando *Exetensions Sets*.

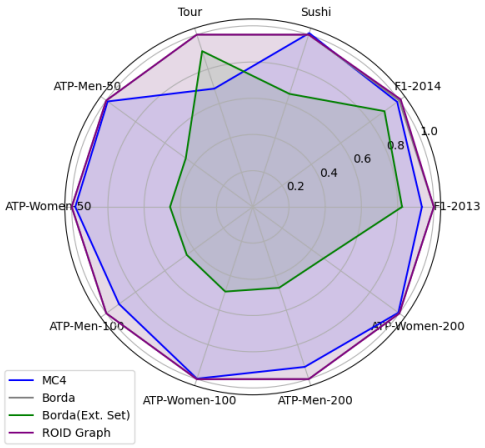

Figure 2: Comparação entre os algoritmos implementad

Isso demonstra, fortemente, o poder da utilização de possíveis combinações adicionais que respeitem o núcleo das opiniões no processo de agregação, mesmo que não tenham sido evidenciadas pelo eleitor.

Utilizando o processo de *Local Kemenization* (LK), entretanto, os resultados tem um avanço enorme, como mostrado pela Figura 3.

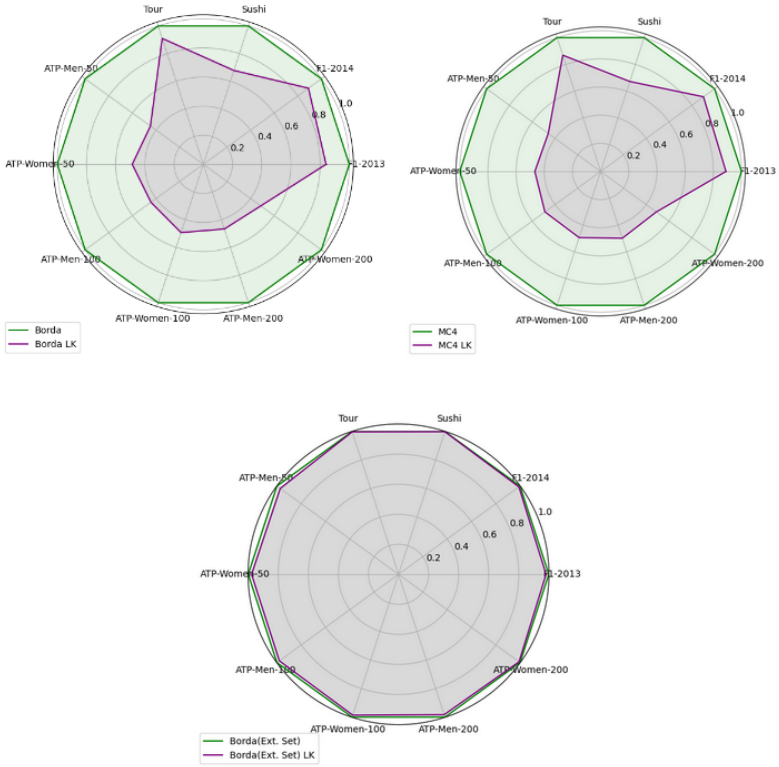

Figure 3: Comparação entre os algoritmos apresentados e os mesmos porém utilizando LK

O que também demostra o poder que a técnica *Local Kemenization* pode ter sobre outras abordagens que não possuem um alto desempenho no processo de agregação de *rankings*.

4 Conclusão

Neste trabalho, investigamos três abordagens distintas para o problema de agregação de rankings, complementadas por uma etapa de refinamento local baseada em trocas adjacentes. A análise comparativa demonstrou que não existe um método universalmente superior, mas sim abordagens mais apropriadas a diferentes contextos e níveis de incompletude.

Para trabalhos futuros, pretende-se investigar combinações híbridas dessas abordagens, bem como utilizar desse conhecimento adquirido sobre Agregação de Rankings para tentar abordar, ou ao menos entender como reacair nessa instância mais específica, o problema de maneira mais geral onde um eleitor não necesariamente possui um ranqueamento conexo, mas sim opiniões esparsas.

References

[1] John G. Kemeny. "Mathematics without numbers". In: *Daedalus* (1959), 88: 571–591.

[2] John G. Kemeny e J. Laurie Snell. *Mathematical Models in the Social Sciences*. Reprinted by MIT Press, Cambridge, 1972. New York: Blais-dell, 1962.

[3] J.J. Bartholdi e C. A. Tovey e M.A. Trick. "Voting schemes for which it can be difficult to tell who won the election". In: *Social Choice and Welfare* (1972), 6(2):157–165.

[4] C. Dwork e R. Kumar e M. Naor e D. Sivakumar. "Rank aggregation methods for the web". In: *Proceedings of the 10th International World Wide Web Conference* (2001), pp. 613–622.

[5] C. Dwork e R. Kumar e M. Naor e D. Sivakumar. "Rank Aggregation Revisited". In: *Compaq Systems Research Center* (2002).

[6] Juan A. Aledo e Jose A. Gámez e David Molina. "Using extension sets to aggregate partial rankings in a flexible setting". In: *Applied Mathematics and Computation* 290 (2016), pp. 208–223. ISSN: 0096-3003.

[7] Yu Xiao e Hong-Zhong Deng e Xin Lu e Jun Wu. "Graph-based rank aggregation method for high-dimensional and partial rankings". In: *Journal of the Operational Research Society* 72.1 (2021), pp. 227–236.