MINERAÇÃO DE DADOS UTILIZANDO DAMICORE: ESTUDO DE CASO DAS ENTREVISTAS DOS ATLETAS OLÍMPICOS BRASILEIROS

Resumo - O Grupo de Estudos Olímpicos, da Universidade de São Paulo, realiza desde os anos 2000 diversas pesquisas no universo olímpico. Uma delas envolve entrevistar os atletas olímpicos brasileiros. Existe uma hipótese no grupo de que, a partir do uso de ferramentas computacionais, novos assuntos e conhecimentos podem surgir a partir da análise dessas entrevistas. Para esse estudo, foi proposta a utilização da ferramenta Damicore, justamente na busca de padrões que podem indicar os caminhos que os pesquisadores podem seguir. Foram realizados diversos testes com o objetivo de se determinar um modelo inicial de análise, saindo de um pressuposto apenas intuitivo para outro, baseado em resultados mais concretos. Foram detectados padrões com relação aos entrevistadores de algumas modalidades, além de apontar quais modalidades provavelmente possuíam assuntos relevantes, dado seu volume de entrevistas.

Palavras-chave: Damicore; Atletas Olímpicos; Mineração de Dados.

DATA MINING USING DAMICORE: CASE STUDY OF BRAZILIAN OLYMPIC ATHLETES 'INTERVIEWS

Abstract – The Olympic Studies Group, from the University of São Paulo, has been conducting research since the 2000s in the Olympic universe. One of them involves interviewing the Brazilian Olympic athletes. There is a hypothesis in the group that, using computational tools, new subjects and knowledge may emerge from the analysis of these interviews. For this study, it was proposed to use the Damicore tool, precisely in the search for patterns that can indicate the paths that researchers can follow. Several tests were performed in order to determine an initial model of analysis, moving from a merely intuitive assumption to another, based on more concrete results. Patterns were detected in relation to the interviewers of some modalities, besides pointing out which modalities probably had relevant subjects, given their volume of interviews.

Keywords: Damicore; Olympic Athletes; Data Mining.

MINERÍA DE DATOS CON DAMICORE: ESTUDIO DE CASO DE LAS ENTREVISTAS DE LOS ATLETAS OLÍMPICOS BRASILEÑOS

Resumen - - El Grupo de Estudios Olímpicos, de la Universidad de São Paulo, ha llevado a cabo varias investigaciones en el universo olímpico desde la década de 2000. Una de ellas, implica entrevistar a atletas olímpicos brasileños. Existe una hipótesis en el grupo que, a partir del uso de herramientas computacionales, pueden surgir nuevos temas y conocimientos del análisis de estas entrevistas. Para este estudio, se propuso utilizar la herramienta Damicore, precisamente en la búsqueda de patrones que puedan indicar los caminos que los investigadores pueden seguir. Se llevaron a cabo varias pruebas con el objetivo de determinar un modelo de análisis inicial, pasando de una suposición que solo es intuitiva a otra, basada en resultados más concretos. Se detectaron patrones con respecto a los entrevistadores de algunas modalidades, además de señalar qué modalidades probablemente tenían temas relevantes, dado su volumen de entrevistas.

Palabras-clave: Damicore; Atletas olímpicos; Minería de datos.

Rovilson de Freitas

Instituto de Ciências Matemática e de Computação

Universidade de São Paulo, Brasil

rovilson.freitas@usp.br

http://dx.doi.org/10.30937/2526-6314.v4.id82

Recebido: 17 dez 2019

Aceito: 24 mar 2020

Publicado: 31 mar 2020

Contexto

Desde o ano de 2001, o Grupo de Estudos Olímpicos da Universidade de São Paulo (GEO/USP), realiza diversas pesquisas no dentro do universo olímpico. Questões como preservação da memória, narrativas biográficas, história de vida, história e evolução do esporte, gênero, racismo, transição de carreira, entre outros, resultaram em diversos artigos, dissertações de Mestrado e teses de Doutorado.

Uma das ferramentas de coleta de dados desse grupo é a entrevista. Lakatos e Marconi[1] definem entrevista como o encontro entre duas pessoas, a fim de que uma delas obtenha informações a respeito de determinado assunto, mediante uma conversação de natureza profissional. No caso do GEO, foram mais de 1100 entrevistas, armazenadas em arquivos texto.

As entrevistas são realizadas com sujeitos que podem, de alguma maneira, contribuir para a pesquisa, trazendo elementos de sua vivência em alguma atividade específica. A experiência pessoal, em determinadas situações, é fundamental para o entendimento do todo sobre aquela atividade. Ninguém melhor para falar sobre alguma coisa do que aquele que viveu ou presenciou o fato.

No caso específico do GEO, as entrevistas têm uma característica especial. Elas não possuem roteiro ou perguntas definidas. Elas partem de uma única pergunta, que permeia a sequência: “Me conte a sua história”. A partir daí, sob a perspectiva da fala do atleta, as informações são coletadas. Perguntas podem surgir, mas sempre a partir da fala desse atleta.

Esse formato de entrevista acaba dificultando a análise dos dados. Por seu caráter totalmente não-estruturado, acaba sendo muito difícil determinar um padrão de respostas por exemplo, visto que não há necessariamente a repetição de perguntas entre os entrevistados. Isso acaba gerando a necessidade do uso de ferramentas computacionais que possam, de maneira automática, apresentar possíveis assuntos a serem estudados pelos pesquisadores. Realizar uma análise manual com as mais de mil entrevistas, além do enorme tempo necessário (e que muitas vezes não disponível num contexto de estudo acadêmico), além da grande possibilidade de dados não serem considerados ou nem mesmo considerados, mostra-se muito difícil e até mesmo, inviável.

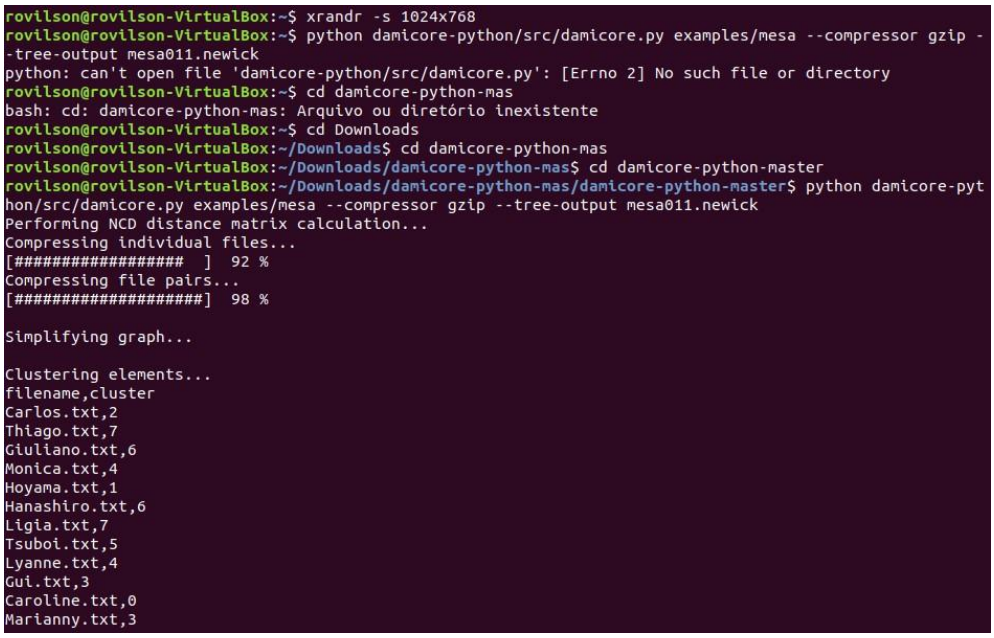
Objetivos do estudo

O principal objetivo desse estudo é, utilizando a ferramenta Damicore, analisar as entrevistas realizadas pelo grupo de estudos olímpicos. A partir dessa análise, verificar os possíveis padrões e agrupamentos resultantes, possibilitando aos pesquisadores uma visão geral das possibilidades dentro desse universo. Importante ressaltar que os resultados obtidos pelo Damicore, apenas, podem ser insuficientes para uma conclusão mais complexa. Entretanto, já apresenta um ponto de partida para os pesquisadores, minimizando assim, tarefas que poderiam prejudicas a pesquisa, como leituras desnecessárias, enfoques incorretos ou abordagens não conclusivas.

Damicore

A metodologia Damicore (Figura 1), desenvolvida por Sanches, Cardoso e Delbem[2], apresenta um método de agrupamento em quaisquer tipos de dado. Isso é possível através da Distância de Compressão Normalizada (NCD), que trabalha como métrica de comparação de elementos, aplicando um compressor genérico sobre a representação binária dos dados, comparando esses elementos comprimidos. Uma de suas características principais é a não necessidade de configuração, não sendo necessário selecionar nenhum parâmetro a priori[3].

Figura 1 - Utilizando o Damicore e Clusters gerados.



Após submeter os dados a análise da ferramenta, um arquivo do tipo newick é gerado. Esse arquivo, é aplicado a uma outra ferramenta chamada Figtree (Figura 2), que representa graficamente esses agrupamentos através de uma árvore. Essa árvore representa a relação de distância entre os objetos analisados[4]. Significa que, quanto mais próximos estão os ramos dessa árvore mais relacionados estão os arquivos analisados (Figura 3).

Figura 2 - Árvore filogênica gerada pelo Figtree.

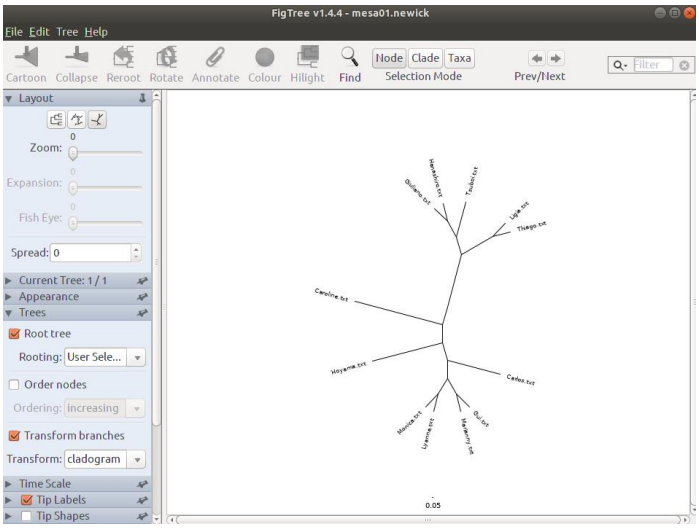
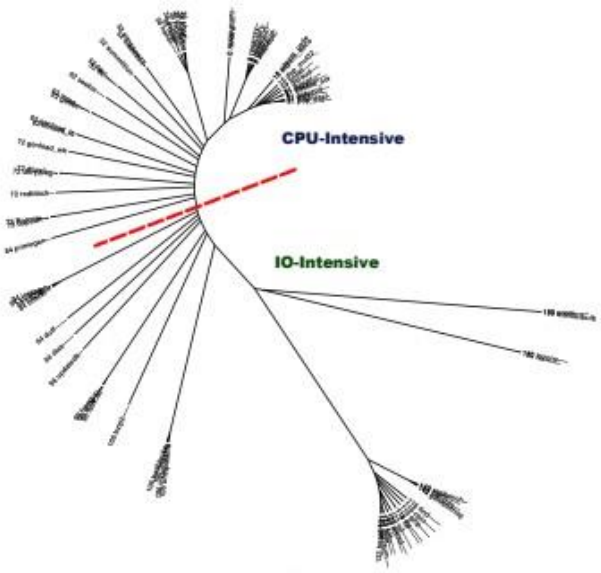


Figura 3- Exemplo de árvore filogênica.



Fonte: Pinto, Delbem e Monaco[5]

Dados selecionados

Para esse trabalho, serão utilizadas as entrevistas dos atletas olímpicos brasileiros, coletadas pelo Grupo de Estudos Olímpicos da Universidade de São Paulo (GEO-USP). Esse grupo, liderado pela Professora Doutora Katia Rubio, trabalha a aproximadamente vinte anos, pesquisando sobre o olimpismo no Brasil. Uma de suas principais linhas de pesquisa vem justamente na preservação da memória dos atletas olímpicos brasileiros.

Lakatos e Marconi[1] definem entrevista como o encontro entre duas pessoas, a fim de que uma delas obtenha informações a respeito de determinado assunto, mediante uma conversação de natureza profissional.

No caso do GEO, as entrevistas têm uma característica especial. Elas não têm um roteiro pré-definido. Todo o processo inicia com uma única pergunta, que direciona os passos seguintes: me conte sua história de vida.

São mais de 1100 entrevistas, gravadas em vídeo e transcritas em texto. E, justamente por sua natureza não-estruturada, nota-se um desafio na procura de relações entre os entrevistados.

Todo trabalho de pesquisa feito até então, era feito de maneira absolutamente manual. Não havia nenhuma ferramenta informatizada para auxiliar nesse processo. Isso resulta em diversos problemas para o pesquisador. O tempo gasto para esse processo, era demasiado alto. Leituras desnecessárias poderiam atrasar ainda mais a publicação de um trabalho. Vários detalhes poderiam não ser detectados pelo pesquisador, considerando todas as características humanas. Entrevistas fundamentais para uma determinada situação poderiam simplesmente ser ignoradas.

Uma ferramenta computacional, além da economia no tempo empregado nas tarefas, poderia apontar padrões dentro de um determinado contexto. Isso auxiliaria o pesquisador a focar sua investigação em determinadas partes do acervo, e não em outras que podem não contribuir nesse momento.

No caso das entrevistas do GEO, são várias as possibilidades de análise. Elas podem partir de questões gerais, comparativas, especificas, entre outras. Para esse estudo, foram realizadas diversas experiências, objetivando perceber padrões através dos agrupamentos apresentados pelo Damicore. Os resultados dessas análises dependem do objetivo naquela situação. Muitas vezes, se mostram desnecessárias ou com nenhum resultado prático. Em outros casos, mostram tendências, que apresentadas ao GEO,

podem proporcionar diversas reflexões. Todas as entrevistas foram convertidas em formato texto (.txt).

Trabalhos Relacionados

O trabalho "Caracterização do perfil de consumo de recursos de programas binários utilizando a técnica Damicore", apresentado por Pinto, Delbem e Monaco[5] no XIII Simpósio Brasileiro de Sistemas de Informação, apresenta o uso da técnica de mineração de dados Damicore para promover a caracterização do perfil de consumo de programas binários. Foi aplicado com sucesso, num conjunto composto por 80 programas binários selecionados.

O trabalho apresentado no Congresso Brasileiro de Informática na Educação, por Moro et al.6 em 2015, intitulado "Caracterização de Alunos em Ambientes de Ensino Online: Estendendo o Uso da Damicore para Minerar Dados Educacionais", traz resultados parciais de uma pesquisa de mestrado que identificou padrões relevantes de comportamento de alunos que interagiram com um sistema educacional web, que poderá resultar no futuro uma ferramenta que auxilie os professores.

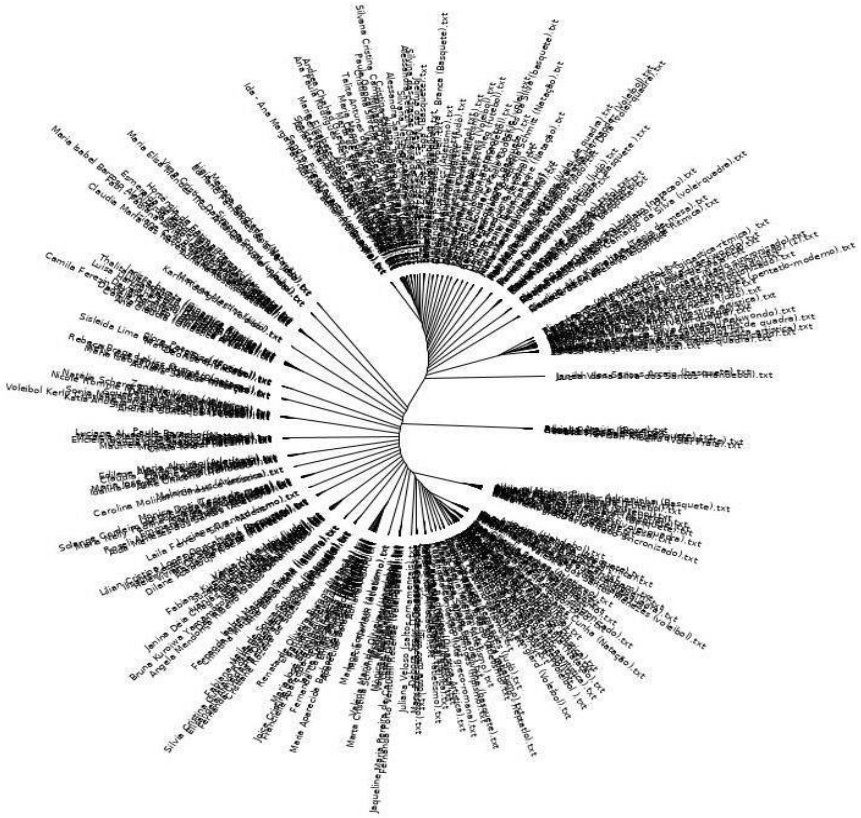
Experimentos realizados e discussão

1. Gênero

A primeira hipótese de análise considerada foi a de gênero. Das entrevistas selecionadas, 777 eram de atletas homens e 375 de atletas mulheres. Deveriam ser verificados os padrões dentro de cada gênero, e eventualmente, os padrões na comparação entre os dois.

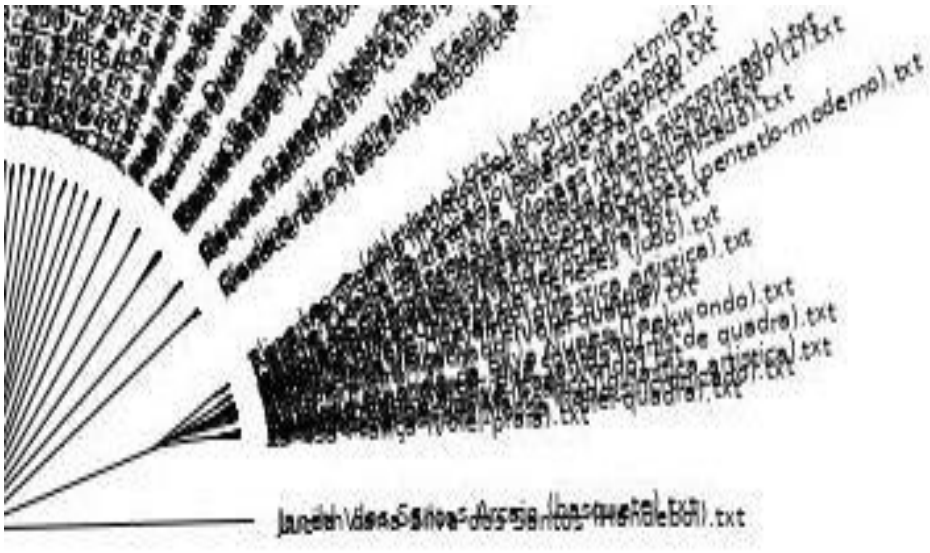
O primeiro resultado (Figura 4), considerando apenas as mulheres, mostrou dois grandes grupos predominantes. Um deles maior, o outro mais concentrado.

Figura 4 - Árvore gerada com as entrevistas das atletas mulheres.



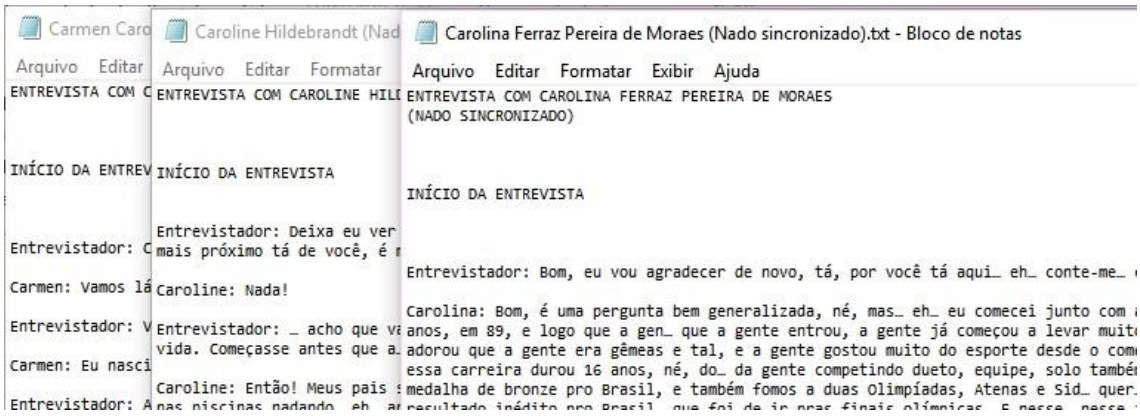
Cada linha da árvore representa uma entrevista. A princípio, não foi percebido nenhum padrão aparente (por modalidade, por exemplo). Um ramo específico foi selecionado, para uma investigação específica (Figura 5).

Figura 5 - Ramo selecionado da árvore das atletas mulheres.



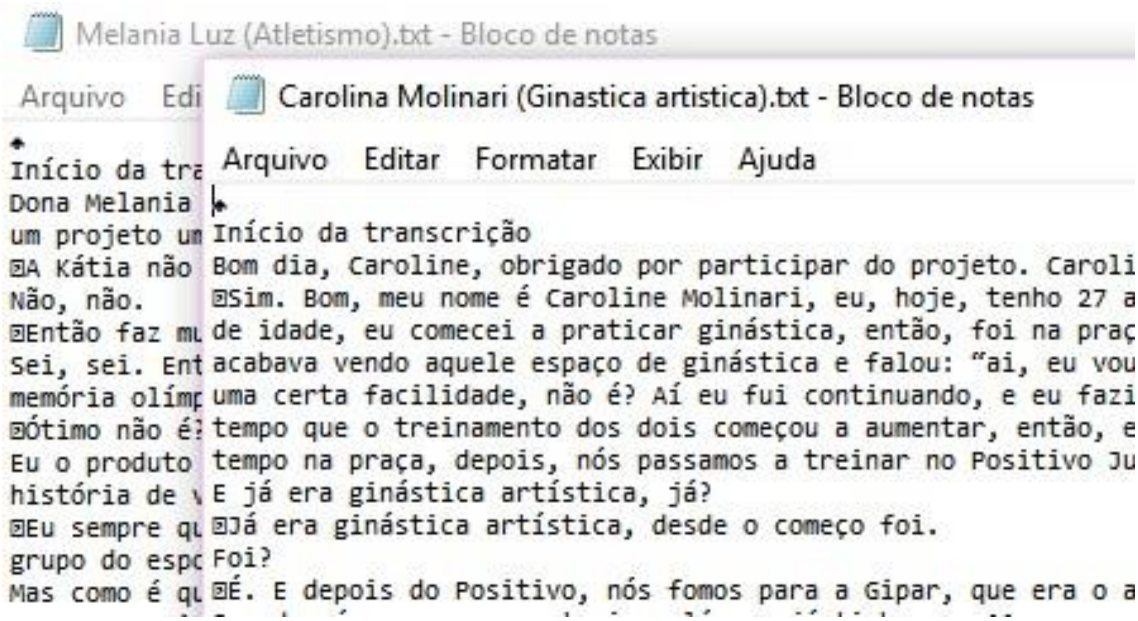
Esse ramo trazia atletas de diferentes modalidades (Judô, Pentatlo, Vôlei etc.). Ao verificar algumas das entrevistas mais próximas apresentadas, notou-se que elas traziam um mesmo padrão de transcrição (Figura 6 - com determinadas palavras, símbolos gráficos, entre outros).

Figura 6 - Padrão de transcrição de entrevistas.



Para se comprovar essa hipótese, um ramo de um lado oposto ao selecionado anteriormente foi escolhido para verificar se o mesmo padrão se confirmava (Figura 7). O resultado foi o mesmo (padrão de transcrição muito parecido).

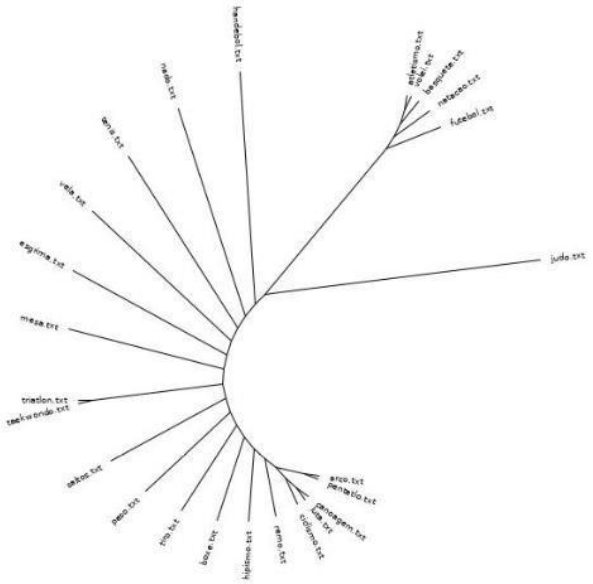
Figura 7 - Padrão de transcrição de entrevistas - ramo oposto.



O mesmo se verificou com os atletas do masculino. Ao aplicar as entrevistas dos homens, também se percebeu um agrupamento pelo padrão de transcrição.

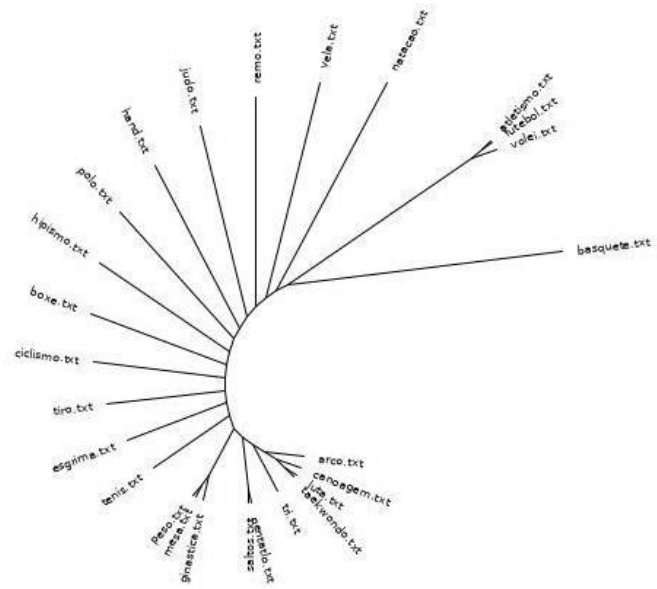
A segunda hipótese dizia respeito ao comparativo ente modalidades e gêneros. As entrevistas foram agrupadas por modalidade, no masculino e no feminino (Figura 8). Foi notada um padrão, tanto no masculino, quanto no feminino, de algumas modalidades mais próximas.

Figura 8 - Modalidades Feminino.



No feminino, as modalidades agrupadas na parte superior foram: Atletismo, Futebol, Vôlei, Basquete, Natação. Essas são as modalidades com o maior número de entrevistadas. Esse agrupamento pode indicar que, essas modalidades em especial, tem potencial de descoberta de conhecimentos, visto que são as que apresentam maior quantidade de informações. Para o masculino, foi realizado o mesmo teste (Figura 9).

Figura 9 - Modalidades Masculino



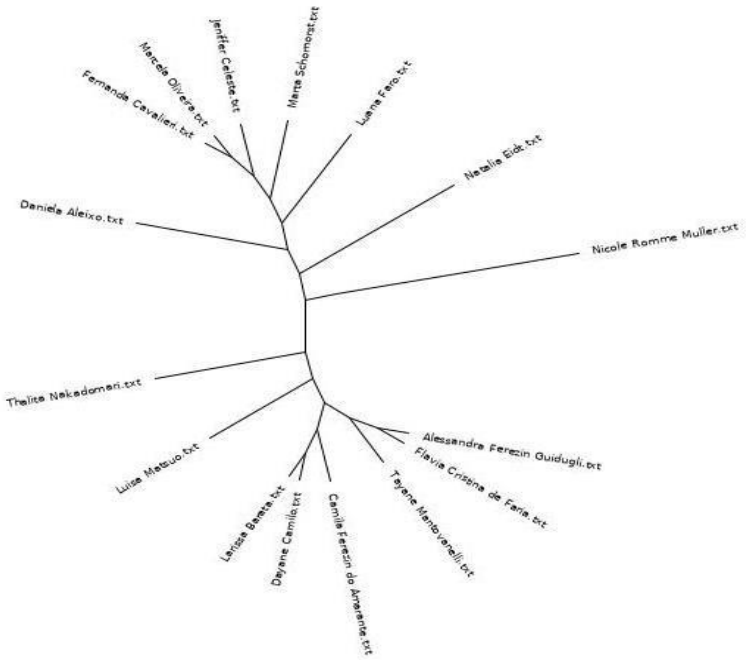
Também houve um agrupamento na parte superior, mas com uma quantidade menor de modalidades. Ainda assim, elas representavam as que tinham o maior número de informações (Atletismo, futebol e vôlei).

2. Considerando apenas uma modalidade

Outra possibilidade testada no Damicore, foi a de que, dentro de uma mesma modalidade, poderia existir algum tipo de padrão. Para esse teste foram consideradas duas modalidades: ginástica rítmica e tênis de mesa. Elas foram selecionadas justamente porque ambas possuem as entrevistas de todos os atletas participantes de uma edição de Jogos Olímpicos até 2012.

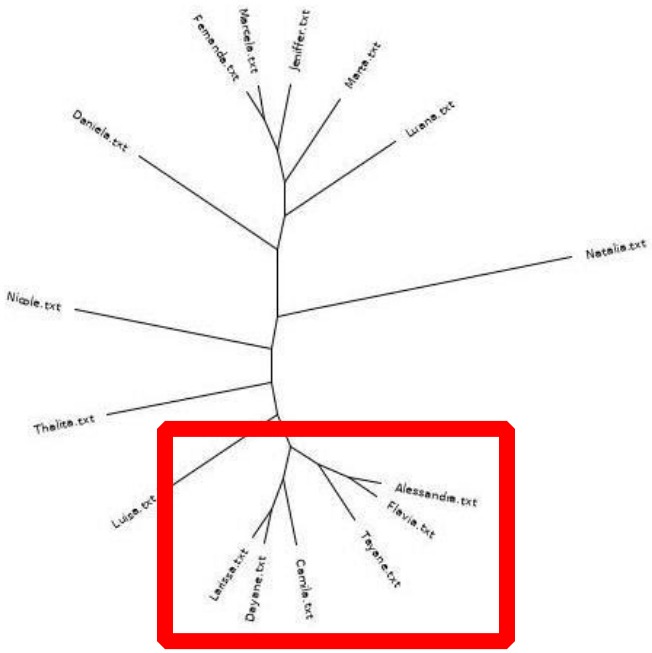
No caso da ginástica rítmica, foram constatados dois grupos importantes (Figura 10). Porém, no geral, o resultado dessa modalidade mostrou uma dispersão maior do que nos casos anteriores. Isso pode ser resultado do número de entrevistas utilizado.

Figura 10 - Entrevistas completas da Ginástica Rítmica.



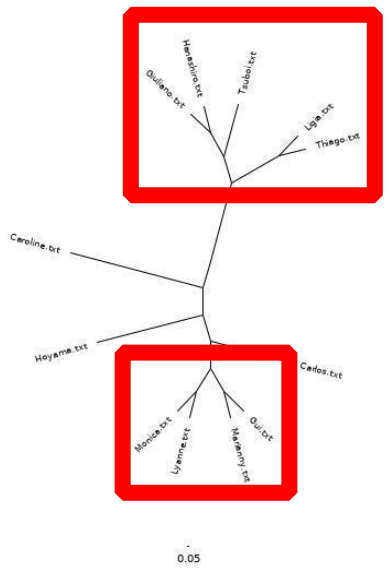
Notou-se que, no grupo mais destacado (parte inferior), as entrevistas foram realizadas pelo mesmo entrevistador (Figura 11).

Figura 11 - Grupo destacado - Mesmo entrevistador



A partir dessa descoberta, foi realizado um teste com o tênis de mesa. E nesse novo teste, foram percebidos agrupamentos de entrevistas realizadas pelo mesmo entrevistador (Figura 12).

Figura 12 - Entrevistas Tênis de Mesa - Destaque para entrevistas com o mesmo entrevistador.



Considerando esse resultado (em que o agrupamento se deu através do entrevistador), o GEO pode, por exemplo, estabelecer estratégias para as futuras entrevistas. Novas abordagens no treinamento dos entrevistadores, novas técnicas empregadas etc.

Considerações Finais

Ao aplicar as entrevistas dos atletas olímpicos brasileiros à metodologia

Damicore, supunha-se que alguns padrões poderiam ser descobertos dentro desse acervo. Após diversos testes, notou-se que realmente alguns padrões surgiram. Novos testes podem indicar novas tendências, considerando que a natureza desses dados aponta possibilidades diversas.

Os testes apontaram as possíveis modalidades com potencial de pesquisas importantes, considerando os seus volumes de dados. Quanto mais entrevistas de uma

determinada modalidades, maiores são as possibilidades. Outra informação importante notada nos testes, é a importância da maneira como as transcrições são realizadas ou convertidas. Uma padronização nesse processo pode ajudar no encontro de dados importantes para os pesquisadores. E finalmente, o papel do entrevistador no processo pode estabelecer alguns padrões de respostas dos entrevistados, ainda que nesse caso em especial as entrevistas não sejam necessariamente estruturadas. Alguns fatores podem contribuir para essa influência, como, por exemplo, entrevistas realizadas no mesmo dia. Com isso, o GEO tem condições de tomar decisões objetivando melhorar esse procedimento.

Entretanto, é importante ressaltar que, apesar de colaborar com o processo, O Damicore não apresentará uma solução definitiva. Ele, combinado com outras técnicas, ferramentas ou metodologias, auxiliam no processo, mas dependem de um trabalho complexo para atingir os objetivos propostos.

Referências

1 Lakatos EM. Marconi MA. Fundamentos de metodologia científica. São Paulo: Atlas; 1999.
2 Sanches A, Cardoso JMP, Delbem ACB. Identifying merge-beneficial software kernels for hardware implementation. International Conference on Reconfigurable Computing and FPGAs; November 30 to December 2, 2011; Cancun, Mexico: 2011. p. 74-79.
3 Cesar BKM. Estudo e extensão da metodologia DAMICORE para tarefas de classificação [dissertação]. São Carlos: Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação; 2016.
4 Valdivia AMC. Mapeamento de dados multidimensionais usando árvores filogenéticas: foco em mapeamento de textos [dissertação]. São Carlos: Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação; 2007.
5 Pinto R, Delbem A, Monaco F. Caracterização do perfil de consumo de recursos de programas binários utilizando a técnica DAMICORE. XIII Simpósio Brasileiro de sistemas de informação; maio 2017; Lavras, Minas Gerais: 2017. p. 128-134.
6 Moro LFS, et al. Caracterização de alunos em ambientes de ensino online: Estendendo o uso da DAMICORE para minerar dados educacionais. 3º Congresso Brasileiro de Informática na Educação; janeiro 2015; Dourados, MS: 2015. p. 631-640.