Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Explorando as relações entre os aspectos de novidades musicais e as preferências pelos ouvintes

Andryw Marques Ramos

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Metodologias e técnicas da computação

Nazareno Ferreira de Andrade
(Orientador)

Campina Grande, Paraíba, Brasil
Andryw Marques Ramos, 05/09/2014

Resumo

A busca por novidades musicais, sejam elas músicas, álbuns ou artistas, é um aspecto central no hábito das pessoas quando se trata de música. E esta procura aumentou principalmente por causa da grande quantidade de música disponível e com fácil acesso proporcionado pelo avanço de tecnologias como Last.FM, Sportify, Youtube, Itunes, entre outros. Porém, devido a esta grande disponibilidade, nem sempre é fácil a descoberta de novidades que sejam relevantes. Para resolver este problema, muitos esforços foram elaborados. O presente trabalho tenta expandir estes esforços tratando a novidade de maneira multidimensional, de acordo com dois aspectos: familiaridade (o quanto o ouvinte conhece outras músicas/ artistas similares à novidade) e popularidade (o quão essa música / artista é conhecida pelos ouvintes em geral). O presente trabalho corrobora esta visão multidimensional da novidade, que é uma visão mais rica e que pode aperfeiçoar ferramentas que dão suporte a descoberta de novidades para ouvintes, como sistemas de recomendação, sites, fóruns, etc. Desta maneira analisamos as preferências dos ouvintes por artistas com novidade (artistas que nunca foram escutados anteriormente pelo ouvinte) baseadas nestes dois aspectos. Para isso foi estudado os hábitos de escuta dos usuário do Last.FM, rede social musical que registra o que os usuários escutam. Os resultados sugerem que não existe uma preferência geral dos ouvintes por algum aspecto das novidade. Os ouvintes tendem a formar grupos baseados nas preferências pelos aspectos das novidades. Estes resultados sugerem um tratamento específico para estes grupos de ouvintes, como um sistema de recomendação que leve em conta estas preferências. Outro estudo realizado neste trabalho compara as preferências dos ouvintes pelos aspectos tanto dos artistas com novidade quanto dos artistas já conhecidos. Este estudo apontou que as preferências dos ouvintes para estes dois âmbitos são diferentes, onde os ouvintes tendem a formar grupos baseados nestas diferenças de preferências. Este resultado implica que o âmbito das novidade e o âmbito do que já se conhece não deve ser tratado da mesma maneira.

Abstract

The search for new music, e.g. songs tracks, albums or artists, is a central aspect in the habit of people when it comes to music. And this pursuit increased mainly because of the large amount of available music and with easy access provided by the advance of technologies like Last.FM, Sportify, Youtube, Itunes. However, due to this high music availability is not always easy to discover relevant novelties. This study attempts to expand the studies about music novelties by investigating how the music preferences of listeners are affected by two different aspects of novel artists: familiarity (how much the listener knows other artists similar to novelty) and popularity (how this artist is known by listeners in general). The study supports this multidimensional view of novelty, which is a richer view and it enables the improvement of tools that support the discovery of music novelties for listeners, as recommender systems, websites, forums, etc.. We collected and analyzed historical data from Last.fm users, a popular online music discovery service. The results suggest that there is not a general preference for some aspect of novelty. Listeners tend to form groups based on the preferences for the novelty aspects. These results suggest a specific treatment for these groups of listeners, e.g. a recommendation system considering these preferences. Another study performed compares the listeners preferences by aspects of both novelty artists and artists already known. This study showed that the listeners preferences for these two spheres are different, where listeners tend to form groups based on these different preferences. This result implies that the scope of novelty and the scope of what is already known should not be treated the same way.

Agradecimentos

Agradeço a todos que me ajudaram até aqui.

Conteúdo

1 Introdução **1**

1.1 Motivação 1

1.2 Perguntas de pesquisa 2

1.3 Resultados 3

1.4 Organização do documento 4

2 Novidades e descobertas no âmbito musical **5**

2.1 Novidade no consumo em geral 5

2.2 Incorporação de novidade em sistemas de recomendação 6

2.3 Grupos de pessoas baseados no comportamento 7

2.4 Nossas contribuições 8

3 Conceitos e Modelos **10**

3.1 Tipos dos itens estudados 10

3.1.1 ~~Item com novidade /~~ Novidade 10

3.1.2 Item conhecido 11

3.2 Características de um item 11

3.2.1 Descritores 11

3.2.2 Descrições 11

3.3 Representações do histórico musical 12

3.3.1 Artista 12

3.3.2 Perfil 12

3.3.3 Ecleticidade 14

3.4 Aspectos 16

3.4.1 Familiaridade 16
3.4.2 Popularidade 17
3.5 Preferência dos itens pelos ouvintes 17

4 Dados utilizados **18**
4.1 Last.FM 18
4.2 Ouvinte 19
4.2.1 Linha do Tempo 19
4.2.2 Histórico do usuário 20
4.2.3 Filtros 21
4.3 Metadados dos Artistas 22
4.4 Aplicação dos dados nos modelos 24
4.4.1 Perfil 24
4.4.2 Ecleticidade 25
4.4.3 Familiaridade 27
4.4.4 Popularidade 27
4.4.5 Preferências 28

5 Preferências dos ouvintes para diferentes aspectos de novidades **30**
5.1 Preferências gerais 30
5.2 Preferências individuais 31

6 Grupos de usuários para diferentes aspectos de novidade **34**
6.1 Conjunto de sujeitos 34
6.2 Dados que caracterizam sujeitos 35
6.3 Algoritmo de agrupamento 36
6.4 Escolha do número de grupos 36
6.5 Grupos 37
6.6 Discussão dos grupos encontrados 40

7 Comparação das novidades com os artistas conhecidos **43**
7.1 Seleção de sujeitos 43
7.2 Características dos artistas conhecidos 44

7.3 Comparação entre relações das preferências e aspectos das novidades e dos artistas conhecidos 44
7.4 Grupos de ouvintes baseados na diferença das relações entre preferências e aspectos das novidades e dos artistas conhecidos 47

8 Conclusão **50**
8.1 Resumo 50
8.2 Implicações 51

Lista de Tabelas

4.1 Sumário dos dados coletados. 22

4.2 Tags do artista Michael Jackson, junto com o valor normalizado de cada uma. 24

4.3 Número de ouvintes (popularidade) de alguns artistas no Last.FM 24

5.1 Correlação (Coeficiente de Spearman) entre aspectos da novidade e preferências, analisando todas as novidades juntas 31

6.1 Centróides para as configurações com 7 grupos e com 8 grupos 39

7.1 Correlações calculadas para artistas com novidade e artistas conhecidos. . . 45

7.2 Teste-T pareado entre as correlações dos aspectos e preferências dos ouvintes para artistas com novidade e artistas conhecidos. 45

7.3 Correlação (Coeficiente de Spearman) entre correlações calculadas para novidades (linhas) e artistas conhecidos (colunas) 46

7.4 Centróides para a configuração com 6 grupos e com 7 grupos 48

Capítulo 1

Introdução

1.1 Motivação

A procura e descoberta de novas músicas e artistas é um aspecto importante no consumo musical. Maddi [19] argumenta que consumidores em geral possuem um *impulso interno*, que tem como finalidade descobrir novas experiências, a fim de criar novos sentimentos e emoções.

A busca por novidades musicais foi alterada com grandes potenciais e desafios nos últimos anos. Serviços de streaming como Spotify, Youtube e Soundcloud, rádios online como a do Last.Fm e até mesmo os sites de compra de música digital, como Itunes e Beatport, possibilitam o acesso a uma grande variedade de músicas. Isso facilita o acesso a músicas e artistas não escutados antes pelo ouvinte, as chamadas novidades. Porém, como há uma grande quantidade de novidades, encontrar aquelas que sejam relevantes acaba sendo uma tarefa custosa.

Para ilustrar, tomemos como exemplo o Itunes. A Itunes Store, loja virtual de músicas do Itunes, possuía em 2012 um acervo com cerca de 26 milhões de músicas ([1]). Escutar músicas aleatoriamente neste catálogo de 26 milhões até encontrar uma novidade relevante é impraticável. Um ouvinte poderia fazer uma pesquisa em ferramentas de busca para encontrar algum site musical, depois ir ao Itunes escutar música por música do que foi pesquisado até encontrar alguma relevante. Ainda assim esta alternativa demanda um bom tempo.

Naturalmente, entender o tipo de música que atrai um consumidor possibilita que criemos ferramentas que o ajudem a encontrar música relevante para um determinado momento. Por

exemplo, sistemas de recomendação podem ser construídos com o intuito de auxiliar os ouvintes nas descobertas de novidades.

Boa parte das soluções que tentam auxiliar os ouvintes nestas descobertas tratam a novidade de forma unidimensional. Por outro lado, neste trabalho partimos do pressuposto de que é possível ver uma novidade sob diferentes dimensões. Podemos caracterizar uma novidade com diferentes aspectos, como a familiaridade (o quanto o ouvinte conhece outras músicas´/ artistas similares à novidade) e a popularidade (o quão essa música / artista é conhecida pelos ouvintes em geral). Um ouvinte pode preferir novidades similares, ou familiares, a músicas que ele costuma escutar, mas preferir novidades não populares, e vice-versa. Por exemplo, um ouvinte que geralmente escute artistas de rap pode gostar de novidades mais desconhecidas de rap, enquanto outro ouvinte que geralmente escute rock gosta mais de novidades populares, ~~independe~~ de gênero musical. Esta visão multidimensional da novidade é ~~mais rica, e sua análise~~ pode tanto melhorar o entendimento dos hábitos dos ouvintes quanto aperfeiçoar ferramentas que dão suporte a descoberta de novidades, como sistemas de recomendação, sites, fóruns, etc.

1.2 Perguntas de pesquisa

Com o intuito de expandir o entendimento sobre o consumo das novidades, conduzimos uma análise sobre o impacto de dois aspectos - familiaridade e popularidade - das novidades para as preferências de ouvintes de música. No nosso estudo, as novidades são artistas que o ouvinte nunca escutou anteriormente. Assim, procuramos responder quatro perguntas de pesquisa.

A primeira pergunta: *Há alguma relação geral entre ~~aspectos~~ das novidades e as preferências dos ouvintes?* Com ~~ela,~~ tentamos descobrir se ~~todos~~ os ouvintes preferem algum aspecto específico da novidade. O objetivo foi encontrar respostas como: "Os ouvintes no geral preferem novidades familiares a seu gosto musical (um ouvinte de rock prefere novidades de rock a novidades de rap)", "Os ouvintes no geral preferem novidades menos populares"ou "não existe uma preferência no geral".

Já a segunda pergunta é uma especificação da primeira: *Individualmente, os usuários preferem algum aspecto das novidades?*. O objetivo foi encontrar respostas como: "75%

dos ouvintes possuem preferência por algum aspecto, sendo que 15% preferem novidades familiares, 30% preferem não-familiares, etc."

Como descobrimos, respondendo à segunda questão, que diferentes usuários preferem diferentes aspectos das novidades, isso nos levou a terceira pergunta: *Quais são os grupos de ouvintes baseados nas preferências pelos aspectos das novidades?* O intuito foi encontrar grupos de ouvintes, baseados nessas preferências pelos aspectos das novidades, junto com algumas características dos hábitos musicais do ouvinte. Encontrar usuários que compartilhem as mesmas características possibilita~~m~~ que ferramentas, como recomendadores, os tratem de forma ~~diferente dos demais~~.

Por fim, a última pergunta compara artistas com novidade e artistas conhecidos: ~~*As relações entre*~~ *as preferências dos ouvintes* ~~*e os*~~ *aspectos das novidades são as mesmas que as* ~~*relações entre as*~~ *preferências dos ouvintes* ~~*e os*~~ *aspectos dos artistas já conhecidos?* O objetivo é verificar se as preferências dos ouvintes pelos aspectos das novidades são semelhantes às preferências dos ouvintes pelos aspectos dos artistas os quais os ouvintes já tinham escutados anteriormente - os artistas conhecidos. Relacionamos estas preferências no mesmo período de tempo para entender se o comportamento dos ouvintes é o mesmo para ambos os tipos de artistas escutados ou se há alguma diferença.

1.3 Resultados

Para responder as perguntas de pesquisa, coletamos dados históricos referentes à escuta de música de usuários do Last.FM, junto com metadados que caracterizam os artistas escutados. Com os dados históricos, conseguimos identificar os artistas com novidade, os artistas conhecidos e as preferências dos ouvintes. Com os metadados, conseguimos identificar os aspectos dos artistas - familiaridade e popularidade.

Descobrimos que, considerando todos os ouvintes de nossa amostra, não há uma correlação entre a familiaridade ou popularidade das novidades e as preferências dos ouvintes. Porém, individualmente, boa parte dos ouvintes preferem um e/ou outro aspecto das novidades.

Como boa parte dos ouvintes preferem algum aspecto da novidade, conseguimos dividí-los em 7 grupos que os distinguem quanto a diferentes preferências: relação entre a famili-

aridade das novidades e as preferências, a relação entre a popularidade das novidades e as preferências, a ecleticidade do ouvinte, a popularidade dos artistas escutados pelo ouvinte e a proporção de novidades que ele escutou no período observado.

Já no âmbito dos artistas com novidade e dos artistas conhecidos, descobrimos que os ouvintes possuem diferentes preferências pelos aspectos da primeira, comparados com os aspectos da segunda. No geral, a preferência por artistas familiares e populares é maior para os artistas conhecidos que para as novidades.

1.4 Organização do documento

O restante deste documento está organizado da seguinte forma. No Capítulo 2, expomos os trabalhos relacionados ao tema da novidade, seja trabalhos que estudam o comportamento das pessoas frente a este aspecto, seja trabalhos que fazem modelos da novidade e que propõem soluções para a recomendação de itens com novidade. Nós relacionamos o impacto do nosso estudo dentro desses trabalhos e apresentamos nossa contribuição. Em seguida, no Capítulo 3, definimos as características das novidades utilizadas na pesquisa. Já o Capítulo 4 descreve quais foram os dados utilizados na pesquisa, enquanto o capítulo 5 descreve os modelos construídos para que o estudo fosse realizado.

A partir do Capítulo 6 são expostos os resultados do trabalho. Nele expomos as relações entre as preferências dos ouvintes com os aspectos das novidades, tanto analisando todos os sujeitos juntos, quanto analisando-os individualmente. O Capítulo 7 utiliza as relações do Capítulo 6 para agrupar os ouvintes. Assim, mostramos quais grupos de ouvintes existem nos dados baseados nas relações entre preferências e aspectos das novidades.

Já o Capítulo 8 expande o estudo com as novidades e compara as relações encontradas sobre preferências e aspectos das novidades com as relações entre preferências e aspectos dos artistas conhecidos. Por fim, retomamos os principais resultados e apresentamos quais as implicações.

Capítulo 2

Novidades e descobertas no âmbito musical

Podemos dividir os trabalhos sobre novidade em três grandes temas: os trabalhos que estudam a faceta da novidade no consumo das pessoas; os trabalhos que retratam soluções de sistemas de recomendação de itens com novidade e trabalhos que tentam identificar grupos de pessoas baseadas no seu comportamento frente a novidades. Neste capítulos mostramos uma visão dos trabalhos destes três temas, explicitando que conceitos e resultados usamos na nossa pesquisa bem como relacionando-os às nossas contribuições. Os trabalhos listados podem envolver estudos sobre novidade tanto no âmbito geral quanto no âmbito musical, como também podem envolver tanto trabalhos sobre comportamento musical frente a novidades como comportamento musical no geral.

2.1 Novidade no consumo em geral

Vários trabalhos estudam o comportamento de novidade e diversidade no consumo das pessoas. Os consumidores em geral podem possuir comportamento com o intuito de manter a consistência de suas escolhas [14] ou buscar a variedade / novidade, com o objetivo de encontrar novos estímulos e evitar a saturação [23]. Este tipo de dualidade no comportamento das pessoas, apesar de parecer contraditória, pode ser vista em vários contextos, como na escolha de um livro, de um CD, ponto turístico para visita, fazer compras na mesma loja ou em lojas diferentes, entre outros [8].

Sobre a busca por novidades no âmbito musical, Laplante [18] constata que adolescentes tendem a buscar novidades junto a pessoas com forte ligação a elas, como amigos e familiares. Essa importância da influência social nas preferências musicais das pessoas foi estudada em outros trabalhos [16; 10], que explicam como a música pode moldar uma identididade social do indivíduo.

Devido a existência desta faceta de varieadade / novidade no consumo das pessoas, construtores de sistemas computacionais, como sistemas de recomendação, estão cada vez mais preocupados em incorporá-la nestes sistemas [13; 26].

2.2 Incorporação de novidade em sistemas de recomendação

Muitos dos algoritmos utilizados em sistemas de recomendação de itens e produtos estão interessados em aumentar a acurácia e precisão dos resultados [13]. O objetivo principal é maximizar a recomendação de itens relevantes ao consumidor. Esses algoritmos, como a filtragem colaborativa [24; 7] e filtragem baseada em conteúdo [22], recomendam principalmente itens parecidos com os que já foram consumidos [13]. Esta forma de recomendação, apesar de aumentar a acurácia dos resultados, pode gerar um problema, já que negligencia o fator da novidade, que faz parte do comportamento das pessoas [8; 28]. Recomendar "mais do mesmo" deixa os usuários do sistema entendiados [13; 29; 28]. Por exemplo, ao pesquisar pelo álbum dos Beatles "White Album" no site Amazon.com, as 10 primeiras recomendações são de outros álbuns dos beatles (Figura 2.1).

Relacionado com essa novidade em sistemas de recomendação, Vargas et al. [26] propõem um modelo que relaciona usuários, itens e novidade. Existe uma diferença entre a descoberta (o usuário conhece um item, que deixa de ser novidade), a relevância (item de interesse do usuário) e a escolha (quando o usuário seleciona um item relevante). Além disso, eles apontam que em geral as soluções envolvendo novidades de itens são apresentadas em dois modelos: o modelo baseado em popularidade e o baseado na similaridade de itens previamente expostos. Este tipo de visão da novidade como dois modelos também é corroborado por Beloggin et. al [3].

O modelo baseado em popularidade define que a popularidade do item está relacionada à

Figura 2.1: Álbuns recomendados pelo site Amazon.com ao pesquisar pelo White Album, dos Beatles. Os 10 primeiros são outros álbuns dos Beatles. Busca realizada em Julho de 2014.

sua descoberta pelas pessoas. Quanto menos popular um item, menos ele foi descoberto pelas pessoas, possuindo uma maior probabilidade de ser novidade para a maioria das mesmas [6]. Vários trabalhos utilizam métricas relacionadas com a popularidade para detectar o quanto os algoritmos estudados expõem novidades [6; 5; 30; 26; 3; 29].

O modelo baseado em similaridade define que há uma maior probabilidade de um item ser novidade para um usuário se este não for similar a outros itens descobertos e escolhidos pelo mesmo. Boa parte das abordagens baseadas neste tipo de modelo agrupam os itens em classes, como uma taxonomia, onde os itens são agrupados / rotulados nas classes a partir da similaridade dos mesmos. Assim, são recomendados para o usuário itens de classes que não são similares a classes anteriormente escolhidas pelo usuário [21; 30; 29].

2.3 Grupos de pessoas baseados no comportamento

Outro tema estudado no âmbito de novidades é a existência de diferentes grupos com diferentes preferências para novidades. Por exemplo, Munson e Resnick [20] descobriram, em

um conjunto de usuários online, subgrupos baseados nas preferências por novas opiniões: apreciadores de diversidade, aversos a desafios e buscadores de apoio. Os apreciadores de diversidade são usuários que se interessam tanto por opiniões similares a suas quanto desafiadoras. Eles não se satisfazem com apenas opiniões similares. Já os aversos a desafios se satisfazem mais com opiniões semelhantes, diminuindo a satisfação se lerem opiniões desafiadoras. Já os buscadores de apoio se satisfazem com um certo número de opiniões semelhantes, que suportam seu ponto de vista, sendo indiferentes a demais opiniões conflitantes. Este resultado mostra que diferentes pessoas possuem diferentes comportamentos frente a novidades (no estudo, novas opiniões).

Sobre a existência de diferentes grupos de pessoas no âmbito musical, Jennings [15] sumariza quatro grupos de pessoas baseados no grau de interesse por música: os eruditos, entusiastas, casuais e indiferentes. Os eruditos são pessoas onde a música é parte principal das suas vidas, possuindo conhecimento musical extensivo; os estusiastas são pessoas que consideram música um aspecto muito importante, mas balanceam o interesse com outros temas; os casuais gostam de música mas possuem consideram outros assuntos bem mais importantes e os indiferentes não se interessam por música.

Já Arhippainen & Hickey [2] conduziram uma pesquisa qualitativa e identificaram ~~cerca~~ de 14 grupos de ouvintes baseados em como eles escutam e usam música no seu dia a dia. Dentre os 14 grupos, podemos destacar os Ouvintes Indie, que buscam constantemente novas músicas, Ouvintes do Mainstream, que escutam tudo que está na mídia e Ouvintes Passivos, que geralmente escutam o que outras pessoas estão escutando.

2.4 Nossas contribuições

Apesar de trabalhos passados utilizarem conceitos de novidade, tanto de itens musicais quanto de itens no geral, não foram encontrados estudos que relacionem diferentes aspectos das novidades com a relevância das mesmas para os usuários. Como apontado na Seção 2.2, a novidade pode ser modelada em no mínimo dois aspectos. Além disso, O' Celma [6] comenta que é importante sabermos a relevância de novidades para os usuários, em um estudo "'centrado no usuário"', para que tenhamos um conhecimento mais completo do seu comportamento. Enquanto algumas abordagens para novidade fazem estudos "'centradas

nos itens"', preocupadas principalmente nas características dos itens (como a popularidade).

Assim, nós unimos os aspectos formalizados por Vagas e corroborados em vários estudos com a ideia de relevância (estudo centrado no usuário), que O' Celma [6] corrobora a importância. Com essa junção de conceitos podemos responder que aspectos das novidades são relevantes para ouvintes e consumidores musicais.

Além disso, outro resultado do nosso trabalho segue a linha de trabalhos sobre grupo de pessoas baseado no comportamento [20; 15; 2]. Descobrimos diferentes grupos de ouvintes baseados nas preferências dos mesmos pelos aspectos das novidades. Mesmo em contextos diferentes do trabalho de Munson[20] (opiniões e música), podemos notar que novidade não pode ser tratado de forma única para todos os indivíduos. Nossos resultados sugerem que há uma necessidade de tratamento específico para cada grupo, em um sistema computacional como sistema de recomendação, por exemplo. O'Celma [4] ~~reforça~~ este tipo de tratamento específico para cada grupo.

Por fim, descobrimos que as preferências dos ouvintes frente aos aspectos dos artistas com novidade são diferentes das preferências frente aos aspectos dos artistas conhecidos. Isso sugere que a novidade no âmbito musical seja estudada especificamente, não podendo ~~generalizar~~ do que o ouvinte já conhece.

Capítulo 3

Conceitos e Modelos

Este capítulo define quais conceitos e modelos foram utilizados na pesquisa. Para a construção dos experimentos, primeiro ~~foi definido~~ o conceito de novidade que iríamos trabalhar junto com o das suas características. Foi importante esta definição inicial pois os termos utilizados na pesquisa (novidade, familiaridade, popularidade, relevância) são termos gerais, que podem possuir mais de um significado, não tendo um consenso da literatura.

3.1 Tipos dos itens estudados

Primeiro ~~foi definido~~ o que seria uma novidade, ou um item com novidade, e consequentemente um item conhecido.

3.1.1 Item com novidade / Novidade

Novidade é o conceito central deste trabalho. Um item com novidade é um item que não foi acessado pela pessoa anteriormente. No âmbito musical, itens podem ser músicas, artistas e álbuns, e as pessoas que escutam esses itens são ouvintes. ~~Mais especificamente,~~ tratamos as novidades como artistas que não foram escutados anteriormente pelo ouvinte. Por exemplo, se em algum momento o ouvinte João escutou o artista Eminem pela primeira vez, Eminem deixou de ser uma novidade para ser um artista conhecido. Antes desse momento ele era considerado uma novidade para João.

3.1.2 Item conhecido

Um item conhecido é ~~o oposto da novidade. Assim, é~~ um item que já foi acessado anteriormente pela pessoa. No nosso trabalho, um artista conhecido é um artista que já foi escutado anteriormente pelo ouvinte. ~~No Capítulo 7 é mostrado o resultado da comparação entre as preferências dos ouvintes pelas características dos artistas com novidade e as preferências do mesmo pelas características dos artistas conhecidos~~.

3.2 Características de um item

Para calcular a familiaridade de um artista para um ouvinte, foi necessário definir características do artista e do perfil musical do ouvinte, já que a familiaridade está relacionada com a similaridade do artista para o perfil do ouvinte. Neste contexto surge os conceitos de descritores e descrição.

3.2.1 Descritores

Um descritor é ~~um~~ um símbolo que descreve ~~/~~ caracteriza um item. No âmbito musical, descritores são termos que podem caracterizar músicas, artistas, álbuns, perfis musicais de ouvintes, etc. Estes descritores podem representar gênero musical (pop, forró), localização (latina, brasileira, britânica), humor (animada, depressiva), entre ~~outros~~.

3.2.2 Descrições

Um item pode ser caracterizado por mais de um descritor, e estes descritores caracterizam-no em menor ou maior grau. Tomemos como exemplo o artista Michael Jackson. Ele pode ser descrito pelos termos Pop, Dance e Soul. Destes termos, Pop o carateriza em um grau maior que Soul. Desta maneira, para uma caraterização completa dos itens, é necessário representar este conjunto de descritores. Para isso usamos o termo descrição. Sejam I o conjunto de itens, D o conjunto de descritores e $g : I \times D \to \mathbb{R}$ a função que denote o grau que o descritor $d \in D$ caracteriza o item $i \in I$. Assim, a descrição θ_i de um item $i \in I$ é definida ~~pela Equação 3.3~~:

$$\theta_i = \{(d_1, g(i, d_1)), \ldots, (d_{|\theta|}, g(i, d_{|\theta|}))\} \tag{3.1}$$

Ou seja, é um conjunto de descritores junto com o grau que cada descritor caracteriza o item. Por exemplo, {(pop, 0,7), (dance, 0,5), (soul, 0,5)} pode ser considerada uma descrição do artista Michael Jackson.

3.3 Representações do histórico musical

3.3.1 Artista

Na nossa pesquisa, os itens com novidade e os itens conhecidos são artistas. Para calcular as métricas que foram utilizadas na pesquisa, definimos um modelo para as características dos artistas. Este modelo do artista possui uma descrição e uma popularidade, que representa o quanto de pessoas no mundo conhecem este artista. Formalmente, sejam A o conjunto de artistas, Θ o conjunto de descrições, $\theta_a \in \Theta$ a descrição do arista $a \in A$ e p a popularidade do artista a, o modelo w que caracteriza a pode ser representado ~~pela Equação 3.2~~:

$$w := (\theta_a, p) \tag{3.2}$$

3.3.2 Perfil

Um perfil de um usuário é um modelo que representa características de um determinado usuário sobre determinado ~~tema.~~ Assim, o perfil musical de um ouvinte é uma representação das músicas ou artistas que ele tipicamente escuta. A necessidade de uma representação do perfil musical do ouvinte surgiu primeiramente para calcular a familiaridade de um artista para o ouvinte. Além disso, utilizamos o perfil para extrair a ecleticidade do ouvinte e para gerar uma representação visual do que foi escutado pelo ouvinte

Uma forma de representação intuitiva do perfil de uma pessoa seria o conjunto de gêneros musicais de artistas que essa pessoa escuta/escutou. Tipicamente, ao perguntar a ouvintes qual seu perfil musical, respostas como estas surgem: "Geralmente escuto artistas de Rock", "Escuto mais bandas de Forró e Pagode". É como se eles abstraíssem os artistas, os agrupando baseado em seus gêneros musicais. Uma representação formal desse conceito se

adequaria à nossa necessidade, pois gêneros musicais, como Forró, Pagode e Rock, são considerados descritores musicais. ~~Dessa maneira seria possível calcular a similaridade entre os descritores de um artista com os descritores do perfil musical do ouvinte.~~

Assim, o perfil musical de um pessoa é formado por um conjunto de pares (descrição, relevância) que ~~descreve~~ para um dado período, as características dos artistas escutados por essa pessoa e a frequência com que artistas com diferentes características foram escutados. Por exemplo, uma pessoa pode ter um perfil composto por {(rock, britânico; 0,2), (rock, brasil; 0,15), (rock, eletrônico; 0,1), ...}, ou seja, ter escutados artistas de rock britânico representando 20% do seu histórico musical, artistas de rock brasileiro representando cerca de 15% do seu histórico e artistas de rock misturado com eletrônico, representando 10% do seu histórico. Formalmente, sejam P_θ o conjunto de descrições do perfil P, $\theta_i \in P_\theta$ uma descrição com relevância r_i. O perfil P de um ouvinte é formado por:

$$\mathrm{P} = \{(\theta_i, r_i)\} \tag{3.3}$$

Construção do Perfil

Uma possibilidade para a construção do perfil é agrupar os artistas de acordo com a semelhança em suas descrições e então extrair uma descrição que represente cada grupo de artista. Para realizar esse agrupamento, foi utilizado um algoritmo de agrupamento hierárquico aglomerativo [12]. Este tipo de algoritmo inicializa cada elemento (em nosso caso cada descrição de cada artista) em um grupo, e a cada passo, ele une os dois grupos mais próximos (similares). Desta maneira, foi necessário definir uma medida de distância, ou dissimilaridade, entre os grupos.

Na maior parte dos métodos utilizados no agrupamento hierárquico aglomerativo, a medida de distância entre grupos pode ser gerada a partir de uma métrica de distância entre os pares de elementos e um critério de união que especifica quais grupos unir em cada passo, em função desta distância. Então, foi definido como métrica de distância entre pares de descrições dos artistas o complemento da similaridade do cosseno entre os vetores de descrições dos artistas e como critério de união (*linkage criterion*) o agrupamento de união pela média (o *average linkage clustering*).

A similaridade do cosseno entre as descrições de dois artistas é definida pela Equação

3.4. Como o algoritmo aglomerativo hierárquico requer uma medida de distância, e não de similaridade, foi calculado o complemento da similaridade do cosseno (Equação 3.5).

$$\cos(\vec{\theta}, \vec{\theta'}) := \frac{\langle \vec{\theta}, \vec{\theta'} \rangle}{\|\vec{\theta}\| \|\vec{\theta'}\|} \tag{3.4}$$

$$\text{dis}(\vec{\theta}, \vec{\theta'}) := 1 - \cos(\vec{\theta}, \vec{\theta'}) \tag{3.5}$$

O *average linkage clustering* [12] é um método de união de grupos baseado na média das distâncias entre cada par de elementos de cada grupo. A distância entre dois grupos é definida pela Equação 3.6. Sejam X e Y grupos, onde $x \in X$ a descrição de um artista do grupo X e $y \in Y$ a descrição de um artista do grupo Y. A distância $d(X, Y)$ entre os grupos X e Y é definida pela média das distâncias de todos os pares $x \in X$ e $y \in Y$ (Equação 3.6).

$$\text{d}(X, Y) := \frac{1}{|X|\,|Y|} \sum_{x \in X} \sum_{y \in Y} dis(x, y) \tag{3.6}$$

Após a definição da distância entre grupos, o algoritmo de agrupamento foi aplicado para os artistas de cada ouvinte separadamente. Como é um método aglomerativo hierárquico, o algoritmo inicia a descrição de cada artista dentro de um grupo separado. Em cada etapa os grupos mais próximos vão sendo aglutinados, até chegar em 1 grupo com todos os artistas. É necessária uma condição de parada para selecionar o número de grupos de um ouvinte adequado. Esta condição de parada vai ser comentada no Capítulo 4.

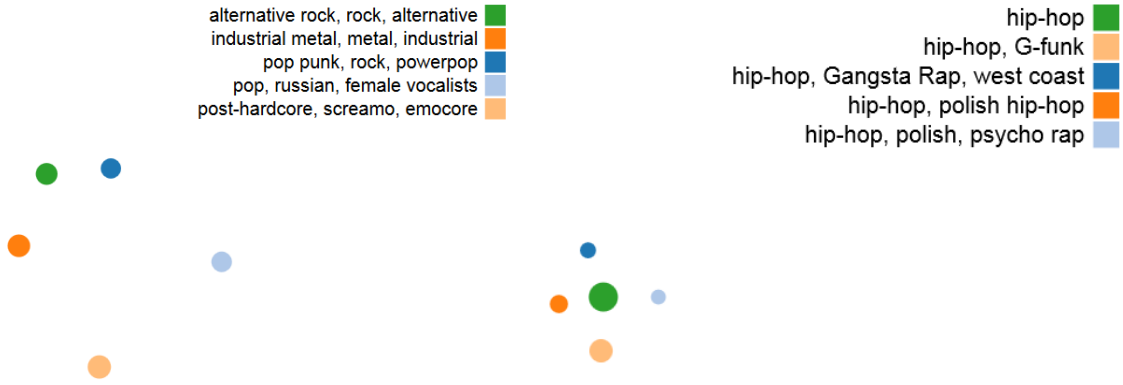
Assim, com o os grupos de artistas definidos, pode-se extrair o conjunto de pares (descrição, relevância) que representa o perfil. Para cada grupo de artistas, o centróide das descrições dos artistas representa a descrição daquele grupo, e a proporção de todas as execuções de músicas dos artistas deste grupo escutadas pelo ouvinte é a relevância deste grupo no perfil do ouvinte. Quanto mais vezes os artistas do grupo i foram escutados, mais influentes as descrições deste grupo são para o ouvinte.

3.3.3 Ecleticidade

A ecleticidade representa o quão eclético musicalmente um ouvinte é - o quão diferente são os grupos de artistas que ele escuta. Ou seja, um ouvinte com alta ecleticidade é um que escuta muitos estilos diferentes de música. Esta métrica foi utilizada para conhecer

melhor os hábitos dos ouvintes e foi utilizada na geração dos grupos de ouvintes baseados nas preferências pelos aspectos das novidades comparadas com seus hábitos musicais, descritos no Capítulo 6.

Inicialmente consideramos utilizar o número de grupos do perfil do ouvinte como critério de ecleticidade. Quanto mais grupos o ouvinte possuísse no perfil, mais eclético ele seria. Porém, dois ouvintes podem possuir o mesmo número de grupos mas um ouvinte pode possuir no perfil grupos mais similares (como um grupo de hip hop e outro de hip hop polonês - Figura 3.1(b)) e outro possuir menos similares (como um grupo de pop e outro de industrial metal - Figura 3.1(a)).



(a) Perfil com alta ecleticidade.

(b) Perfil com baixa ecleticidade

Figura 3.1: Exemplos de dois perfis de ouvintes. Cada círculo representa um grupo de artistas e a distância entre os círculos é proporcional a similaridade entre os grupos. O tamanho de cada círculo é proporcional a quantidade de músicas dos artistas de cada grupo escutadas pelo ~~ouvinte.~~

Uma alternativa a essa abordagem seria contabilizar o quanto de diferença ou diversidade cada grupo adiciona ao perfil. Quanto mais diversidade houver nos grupos do perfil, mais eclético o ouvinte é. Para isso, recorremos novamente a um algoritmo de agrupamento hierárquico, porém agora utilizando as descrições do perfil como elementos a serem agrupados. A cada passo calculamos e armazenamos a distância entre os dois grupos que foram unidos. Por fim, definimos a ecleticidade como a soma de todas estas distâncias. Formalmente, sejam P_θ o conjunto de descrições do perfil do ouvinte; $X^{(s)} := \{X_1^{(s)}, ..., X_n^{(s)}\}$ o conjunto de grupos no passo s (*step* é passo em inglês) do algoritmo hierárquico, onde $X^{(1)} = P_\theta$ e $X^{(j)}$, onde $j > 1$, são grupos do algoritmo hierárquico criados a partir do conjunto inicial $X^{(1)}$.

Seja $d(X_k^{(s)}, X_l^{(s)})$ a distância entre os grupos $X_k^{(s)}$ e $X_l^{(s)}$. Então, a ecleticidade do ouvinte com perfil P é calculada na Equação 3.7.

$$\mathrm{e}(P) = \sum_{s=1}^{|P|-1} min_{k,l \in (1..|P-s+1|)} d(X_k^{(s)}, X_l^{(s)}) \tag{3.7}$$

A distância $d(X_k^{(s)}, X_l^{(s)})$ entre os grupos do algoritmo hierárquico foi calculada utilizando o *average linkage method* (Equação 3.6). Como o *average linkage method* depende da distância entre cada par de elemento (onde cada elemento é uma descrição), foi definido como distância entre dois grupos o complemento da similaridade do cosseno (Equação 3.5) entre as descrições.

3.4 Aspectos

Para caracterizar as novidades multidimensionalmente, utilizamos dois aspectos: a familiaridade e a popularidade. Esta seção define estes dois aspectos.

3.4.1 Familiaridade

Muitos trabalhos [21; 30; 29] caracterizam uma novidade baseada na similaridade do item acessado pela pessoa em relação a outros itens acessados anteriormente pela mesma. Trazendo para o âmbito musical, rotulamos este tipo de característica como familiaridade. A familiaridade de um artista para um ouvinte reflete o quanto este ouvinte foi exposto a outros artistas que têm descrições semelhantes aos do artista escutado.

Além da similaridade das descrições, nós levamos em conta o quanto estes artistas similares ao artista em questão foram escutados pelo ouvinte. Isso porque a familiaridade de um artista para um ouvinte é influenciado também pelo quanto este ouvinte escutou artistas similares ao artista em questão [11]. Por exemplo, se Maria tem hábito de escutar muitos artistas pop e poucos artistas de rock, Britney Spears (cantora pop) é mais familiar a ela que Evanescence (banda de rock). Assim, a familiaridade está relacionada com a similaridade entre a descrição do artista e as descrições do perfil do ouvinte junto com a influência desses artistas no perfil do ouvinte.

Formalmente, sejam P o perfil do usuário, formado pelos pares de (descrição θ_i, relevância r_i) e θ_a a descrição do artista a. Assim, a familiaridade de a para o perfil P do ouvinte é: ~~(Equação 3.8).~~

$$\text{fam}(a, P) = max_{\theta_i, r_i \in P} cos(\vec{\theta_a}, \vec{\theta_i}) \times r_i \tag{3.8}$$

3.4.2 Popularidade

Outra característica ~~bem~~ relacionada com itens com novidade na literatura é a popularidade [6; 5; 30; 26; 3; 29]. A relação feita é que, quanto menos popular um item, menos ele foi descoberto pelas pessoas, possuindo uma maior probabilidade de ser novidade para a maioria das mesmas [6]. Neste trabalho definimos a popularidade como sendo ~~o quanto de pessoas~~ já escutaram o artista em questão. Por exemplo, Michael Jackson, artista que muitas pessoas de todo o mundo já escutaram, é mais popular que Rapadura, um rapper brasileiro que foi escutado apenas por um nicho específico de pessoas. A popularidade faz parte do modelo de um artista, descrito na Subseção 3.3.1.

3.5 Preferência dos itens pelos ouvintes

Com os aspectos familiaridade e popularidade, estudamos quais destas características das novidades são relevantes para o ouvinte. Em outras palavras, qual a preferência dos ouvintes por esses aspectos. Utilizamos dois conceitos de preferência:

1. Atenção total

   É o quanto de atenção que um ouvinte deu para o artista em um período especificado. Para isso utilizamos a quantidade de músicas do artista que o ouvinte escutou no período. Quanto mais músicas do artista, mais atenção o ouvinte devotou ao artista.

2. Período de atenção

   É o período em que o ouvinte devotou de atenção ao artista. No nosso trabalho, a unidade de tempo é uma semana. Assim, quanto mais semanas o ouvinte escutou alguma música do artista em questão, maior o período de atenção do ouvinte.

Capítulo 4

Dados utilizados

Após a descrição das características das novidades (e dos artistas conhecidos) utilizadas no nosso estudo, esta seção descreve os dados que foram utilizados na pesquisa. Podemos dividir os dados em 2 partes: a primeira é representada pelo histórico musical dos sujeitos a serem analisados e a segunda pelos metadados dos artistas escutados pelos sujeitos. Os sujeitos dos experimentos representam os ouvintes. O histórico musical foi utilizado para identificar os artistas com novidade, os artistas conhecidos e as preferências dos ouvintes por ambas. Já os metadados foram utilizados para identificar os aspectos dos artistas - a familiaridade e a popularidade. Os dados foram coletados da plataforma do Last.FM.

4.1 Last.FM

O Last.FM é uma rede social musical que tem como principal característica o *Scrobbling* - um serviço que permite registar o histórico de músicas escutadas pelos usuários. Além disso, o site fornece outros recursos como: serviço de rádio online, recomendador de novidades, tabelas com detalhes do histórico de execução do usuário, informações sobre artistas, turnês e possibilidade de criação de fóruns.

O Last.FM fornece uma API [1](Application Programming Interface - conjunto de rotinas fornecidas por um software para que aplicativos acessem suas funcionalidades) que permite o acesso a dados presentes no site. É possível coletar informações dos usuários, histórico de escuta dos usuários e informações sobre músicas / álbuns / artistas. Para nossos experimen-

[1] www.lastfm.com.br/api

tos, nós coletamos 2 tipos de dados: o primeiro consiste num conjunto de usuários junto com seu histórico de escuta e o segundo em metadados dos artistas escutados. Os usuários do Last.FM foram os sujeitos da pequisa, e como dito anteriormente, representam os ouvintes.

4.2 Ouvinte

Com o intuito de estudar os artistas escutados pelos ouvintes, foram coletados dados acerca de uma amostra de usuários do Last.FM. A amostragem que constituiu esse conjunto de usuários foi feita a partir do procedimento de *SnowBall Sampling* [9], iniciada pelo perfil do autor e sendo expandida pela coleta dos vizinhos musicais. Vizinho musical é um conceito utilizado no Last.FM, onde duas pessoas são vizinhas se possuírem gostos musicais parecidos. Foi coletado um conjunto de 100 mil usuários.

Após a seleção dos usuários, o próximo passo foi coletar o histórico de escuta dos mesmos. Para identificar as novidades, o histórico do usuário foi dividido em períodos, que serão detalhados a seguir.

4.2.1 Linha do Tempo

Os dados referentes ao histórico de cada sujeito foram coletados no período entre a primeira vez que o usuário escutou alguma música no Last.FM e Agosto de 2013. Este período foi dividido em duas partes, como pode-se ver na Figura 4.1: *Histórico Inicial* do sujeito e o *Período de Experimento*. O Histórico Inicial contempla o período desde a primeira música que o sujeito escutou no Last.FM até Agosto de 2012, enquanto o Período de Experimento engloba o período entre Agosto de 2012 e Agosto de 2013 (um ano no total). Além dessa divisão, especificamos os seis primeiros meses do Período de Experimento como *Período de Observação*.

Com esta divisão, foram identificadas quais as novidades escutadas pelo usuário. Os artistas escutados pelo usuário no Período de Observação que não foram escutados no Histórico Inicial são consideradas novidades. Já os artistas que foram previamente escutados no Histórico Inicial são considerados como artistas conhecidos. Não consideramos o Período de Experimento todo para evitar viés no cálculo das características das novidades. Uma novidade a que um sujeito foi exposto no começo do Período de Experimento tem maior

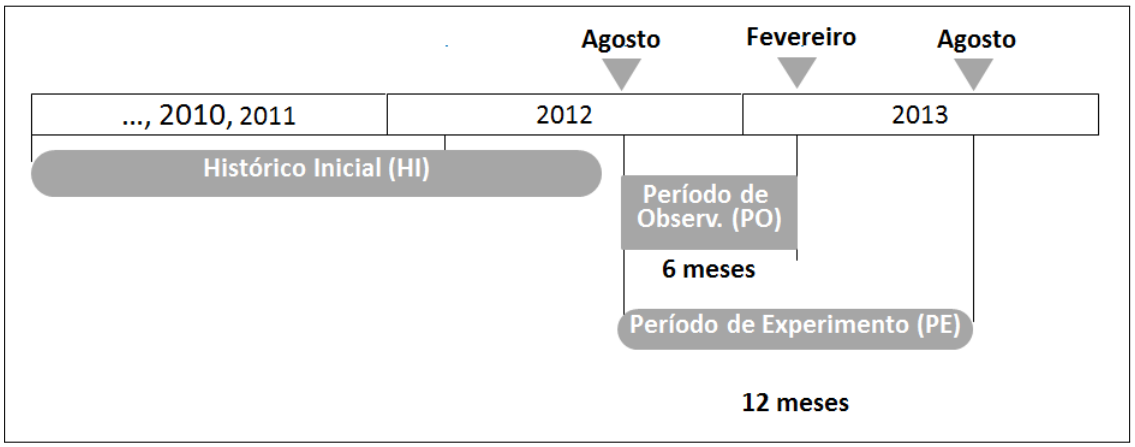


Figura 4.1: Linha do tempo utilizada no trabalho

probabilidade de ser escutada mais vezes que uma novidade à qual o sujeito foi exposto no final do Período de Experimento. Assim, identificamos como novidades os artistas escutados no Período de Observação e levamos em conta as métricas referente a elas durante todo o Período de Experimento (idem para os artistas conhecidos). Isso faz com que nossos dados tenham ao menos 6 meses de observação de cada novidade descoberta pelos sujeitos.

4.2.2 Histórico do usuário

Do Histórico Total do ouvinte, coletamos todos os artistas que ele escutou desde a entrada do usuário no Last.FM até Agosto de 2013, junto com o total de execuções das músicas do artista. O método da API utilizado foi *getTopArtists*, que possibilta coletar os n artistas mais escutados pelo ouvinte, onde este n pode ser até o número total de artistas escutados por ele. Para o Período de Experimento, fizemos dois tipos de coleta. A primeira, utilizando o *getTopArtists* dos 12 meses, coletamos todos os artistas escutados, junto com o número de execuções das músicas de cada. O intuito dessa primeira coleta foi identificar os artistas do Histórico Inicial. A segunda parte consiste nos artistas escutados em cada semana deste período, junto com o número de execuções em cada semana. O método da API do Last.FM utilizado foi o *getWeeklyArtistChart*. Esta segunda coleta foi realizada com o intuito de obter, além do número de execuções das músicas de cada artista, o número de semanas em que o usuário escutou cada artista.

Com os dados do Histórico Total e do Período de Experimento conseguimos identificar o

Histórico Inicial e as novidades. Os artistas do Histórico Inicial são os artistas que o ouvinte não escutou apenas no Período de Experimento. Ou seja, artistas com número de execuções no Histórico Total do ouvinte maior que no Período do Experimento. Já as novidades são os artistas com o mesmo número de execuções no Histórico Total e no Período de Experimento.

Após a coleta e definição de cada período, foi realizado uma filtragem nos dados, descrita a seguir.

4.2.3 Filtros

Como o Last.FM é uma rede social formada por diferentes tipos de usuários com diferentes hábitos musicais, julgamos necessária uma filtragem nos sujeitos, para selecionar os adequados aos propósitos dos experimentos. Abaixo estão as características que os sujeitos precisavam ter para serem selecionados, junto com a maneira de filtragem utilizada.

1. Possuir alta atividade de escuta no período de Histórico Inicial.

   Filtro: Exclusão de usuários que tenham escutado menos de 100 artistas no período de Histórico Inicial.

2. Possuir alta atividade de escuta no Período de Experimento.

   Filtro: Exclusão de usuários que escutaram menos de 100 músicas por semana em pelo menos 1/4 das semanas do Período de Experimento.

3. Serem expostos a um número de novidades que permita a investigação de relações entre as características das novidades e as suas preferências.

   Filtro: Exclusão de usuários que escutaram menos de 10 novidades.

4. Possuir número realista de execuções: foi detectado que alguns usuários possuíam um número muito grande de execuções musicais. Alguns, por exemplo, tiveram uma média de mais de uma música por minuto durante o período de observação, o que na realidade é impraticável. Uma explicação para esse fato seria a criação de robôs que trocassem a música assim que o sistema contabilizasse a execução da música (dependendo da configuração, o Last.FM pode considerar que a música foi escutada se ela foi tocada por um certo tempo, como 30 segundos).

Filtro: Exclusão de usuários que tiveram uma média de execuções superior a 16 horas de execuções por dia, no Período de Experimento. Como as pessoas dormem em média 8 horas por dia, um ouvinte que passe o dia todo, enquanto acordado, escutando música, ele escutaria 16 horas de música por dia. Suponto que uma música tem em média 4 minutos, foram excluídos os usuários que tiveram média maior que 240 músicas por dia ($\frac{16hrs \times 60min}{4min/musica} = 240musicas/dia$*)*

5. Não utilizem majoritariamente a rádio do Last.FM: um dos objetivos da pesquisa é identificar as preferências dos usuários. Assim, é importante que a maior parte dos artistas escutados pelo usuário sejam escolhidos por ele e não por uma rádio.

   Filtro: Exclusão de usuários que não escutaram nenhum artista mais de 15x na semana, em mais de 1/4 das semanas do período de observação. 15 músicas é o número médio de faixas que um álbum contém. Desta maneira, se um usuário escutou mais de 15 músicas de um artista em uma semana, estamos supondo que ou o usuário escolheu escutar um álbum do artista ou selecionou explicitamente 15 músicas do artista. Pela maneira como funciona a licença de direitos autorais para rádios, é bastante improvável que mais de 15 músicas de um artista sejam executadas pela rádio do Last.FM em uma semana.

O processo de filtragem resultou em uma amostra de 11.989 sujeitos. A Tabela 4.1 traz um sumário dos dados coletados.

	Artistas com novidade	Artistas conhecidos
Total	389.853	1.202.869
Média por ouvinte (desvio padrão)	32,5 (26,75)	100,33 (60,03)

Tabela 4.1: Sumário dos dados coletados.

4.3 Metadados dos Artistas

Com o intuito de construir o modelo de artista descrito na Subseção 3.3.1 e com isso calcular os aspectos das novidades, foram coletados dois tipos de metadados referentes aos artistas:

as *tags* que descrevem o artista (representando os descritores do artista) e a popularidade do artista no Last.FM.

Tags são palavras (ou conjunto de palavras), como *rock*, *rap* e *pop*, associadas a um recurso, como músicas, álbuns e artistas . No Last.FM os usuários podem marcar cada um dos recursos com alguma tag, caracterizando-as *tags sociais*. Estas tags podem representar gêneros musicais (rock, samba), localização (brasil, nordeste, germany, west coast), humor (sad, chill, happy), opinião (love, favorite), referência pessoal (seen live, i own it), entre outros. Como as tags podem ser de vários tipos (não apenas gênero musical), elas podem ser consideradas *descritores* dos artistas.

Para cada artista foram coletadas as tags atribuídas a ele pelos usuários, junto com a popularidade de cada tag. Esta popularidade está relacionada à quantidade de vezes que a tag foi atribuída para o artista específico, pelos usuários do Last.FM. A popularidade da tag fornecida pelo Last.FM é normalizada, onde a tag mais atribuída possui valor igual a 100 e as outras tags possuem valores proporcionais, de acordo com a frequência de atribuição de cada uma. Esta popularidade representa o grau de caracterização do descritor para o artista, como discutido na Subseção 3.2.2. Formalmente, sejam A o conjunto de artistas, T o conjunto de tags e $h : A \times T \rightarrow \mathbb{R}$ uma função que denote a frequência absoluta que uma tag $t \in T$ foi atribuída a um artista $a \in A$. O valor normalizado da tag t, representado pela função $f : A \times T \rightarrow \mathbb{R}$ é representado pela Equação 4.1.

$$f(t,a) = \frac{h(t,a)}{max_{x \in T} h(x,a)} \times 100 \tag{4.1}$$

A Tabela 4.2 apresenta as 5 tags com maior valor do artista Michael Jackson. Pode-se ver que a tag *pop* foi a mais atribuída para Michael Jackson, possuindo valor 100. O método utilizado da API do Last.fm foi o *artist.gettoptags*.

Como as tags são associadas pelos usuários do Last.FM, há problemas relacionadas a esse processo [17]. Usuários podem atribuir tags que não condizem com a realidade, podem errar na escrita da tag, etc. Para utilizar tags que realmente descrevam o artista, foi realizado um processo de filtragem. De cada artista foram consideradas as tags populares até um máximo de 4 tags, tendo cada uma popularidade mínima de 30 (onde a tag mais atribuída àquele artista possui valor de 100). Além disso, foram eliminadas manualmente as tags com conotação pessoal, como *seen live* (vi ao vivo) ou *favorite* (favorito).

Tag	Valor
pop	100
80s	49
dance	40
soul	35
funk	32

Tabela 4.2: Tags do artista Michael Jackson, junto com o valor normalizado de cada uma.

Sobre a popularidade do artista, foram coletados o número de usuários do Last.FM que escutaram cada artista. A Tabela 4.3 mostra exemplos de popularidade de alguns artistas. O método utilizado da API foi o *artist.getinfo*.

Artista	Número de ouvintes (popularidade)
Michael Jackson	2.998.428
The Beatles	3.177.625
Red Hot Chili Peppers	4.032.453
Eminem	3.756.890
Chico Buarque	314.584

Tabela 4.3: Número de ouvintes (popularidade) de alguns artistas no Last.FM

4.4 Aplicação dos dados nos modelos

Esta Seção remete os modelos discutidos no Capítulo 3, agora relacioando os modelos com os dados coletados.

4.4.1 Perfil

Como dito anteriormente, o perfil musical de um pessoa é formado por um conjunto de pares (descrição, relevância) que descrevem as características dos artistas escutados por essa pessoa e a frequência com que artistas com diferentes características foram escutados. Para

a construção do perfil foram utilizados os artistas do histórico musical do sujeito que não foram consideradas novidades. Assim, as descrições utilizadas são as tags dos artistas junto com a popularidade de cada tag para o artista.

O método utilizado para construir o perfil foi algoritmo de agrupamento hierárquico aplicado nas descrições dos artistas. Como é um método aglomerativo hierárquico, o algoritmo inicia cada descrição dos artistas dentro de um grupo separado. Em cada etapa os grupos mais próximos vão sendo aglutinados, até chegar em 1 grupo com todos os artistas. Para selecionar o número de grupos de um ouvinte, o algoritmo foi interrompido no momento em que a distância mínima entre 2 grupos fosse igual a 0,30 (lembrando que foi utilizado o complemento da similaridade do cosseno entre os vetores de descrições como cálculo de distância. Os detalhes foram expostos na Subseção 3.3.2). Este valor de 0,30 foi obtido empiricamente. Para tanto, foram selecionados alguns ouvintes com perfis musicais diferentes e foram analisados os grupos criados ao mudar este valor limite. O valor de 0,30 foi melhor valor encontrado, onde - no julgamento do autor desta dissertação e de colegas do mesmo grupo de pesquisa - artistas similares estavam no mesmo grupo e artistas bastante diferentes estavam em grupos diferentes.

Os perfis obtidos tiveram média de 33,28 pares de (descrição , relevância), onde cada par estava relacionado a pelo menos 2 artistas (foram excluídos os pares relacionados a apenas um artista, considerados outliers), e desvio padrão de 20,3. A Figura 4.2 representa o gráfico de distribuição acumulada do número de pares encontrados nos perfis dos sujeitos selecionados em nosso experimento. A Figura 3.1 representa dois exemplos de perfis.

4.4.2 Ecleticidade

Com as descrições dos perfis dos ouvintes calculadas foi possível calcular a ecleticidade de cada ouvinte. A Figura 3.1 compara dois perfis de ouvintes com ecleticidades bastante diferentes. A Figura 3.1(a) é o perfil de um ouvinte mais eclético que o ouvinte da Figura 3.1(b).

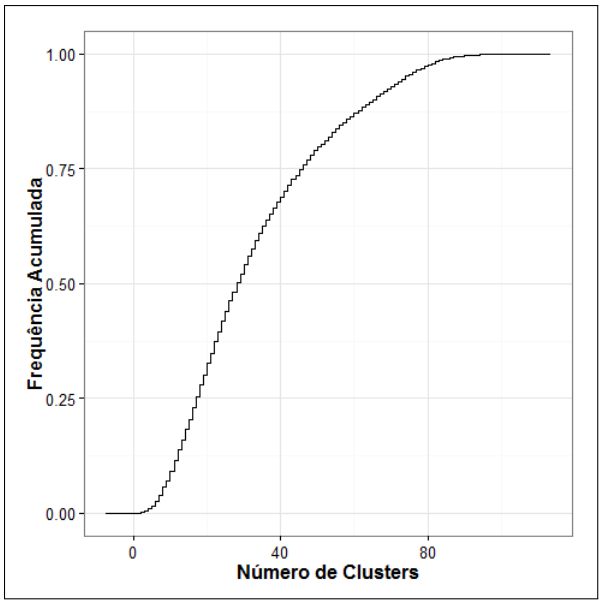
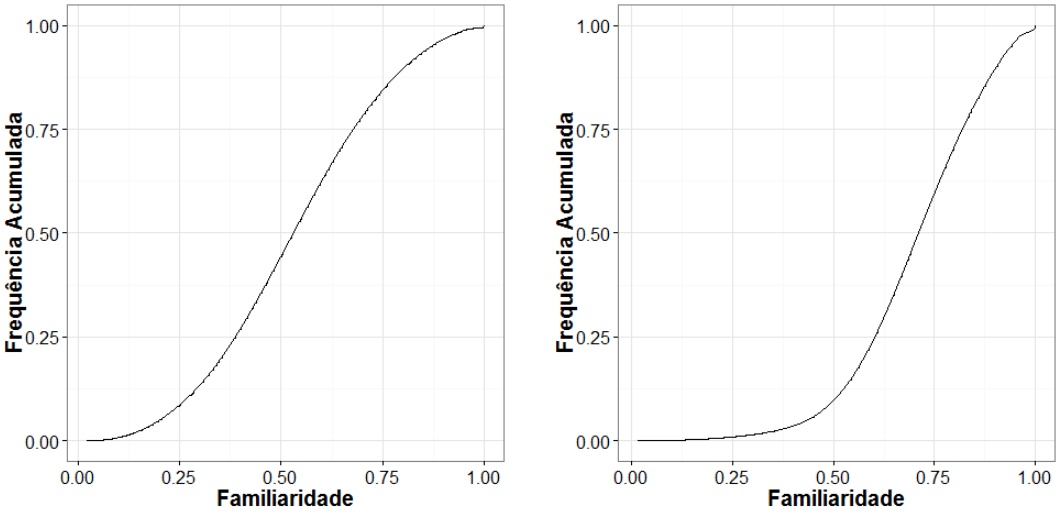


Figura 4.2: Distribuição acumulada do número de grupos



(a) Artistas com novidade

(b) Artistas conhecidos

Figura 4.3: Frequência acumulada da familiaridade dos artistas para os ouvintes.

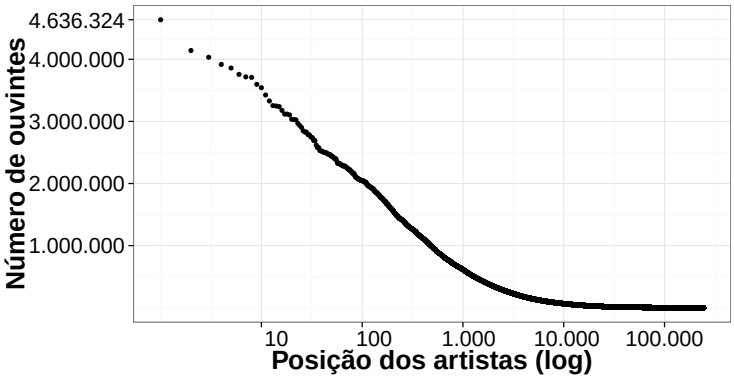
4.4.3 Familiaridade

Com o perfil dos ouvintes calculados foi possível calcular a familiaridade dos artistas para os ouvintes. A Figura 4.3 representa gráficos da distribuição acumulada da familiaridade dos artistas para os ouvintes. Podemos observar que os valores da familiaridade para os artistas com novidade (Figura 4.3(a)) são em geral menores que os valores para os artistas conhecidos (Figura 4.3(b)).

4.4.4 Popularidade

Para calcular a popularidade, utilizamos o logaritmo na base 10 do número de ouvintes do artista no Last.FM. O logaritmo foi utilizado pois a distribuição da popularidade dos artistas é enviesada (4.4).

(a) Número de ouvintes dos artistas

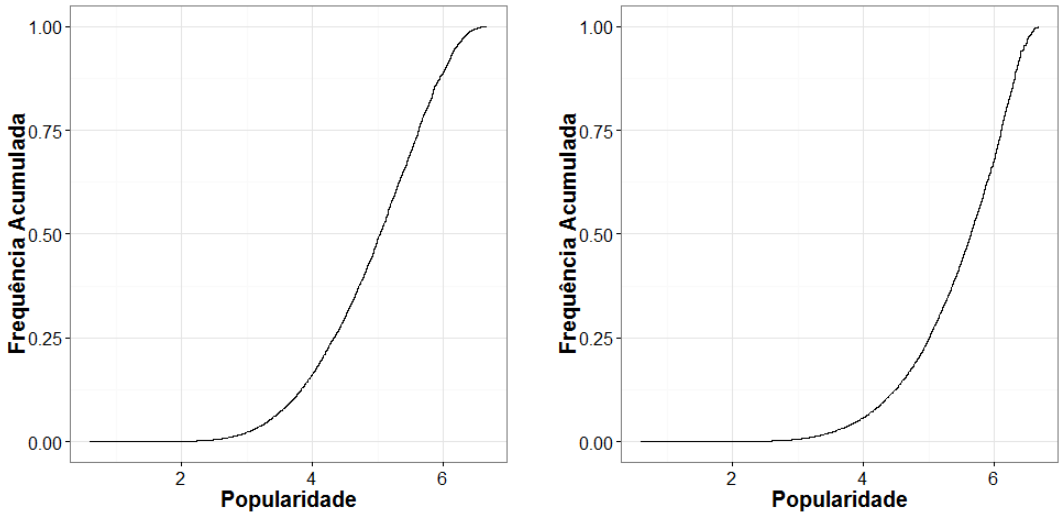


(b) Número de ouvintes dos artistas (log linear)

Figura 4.4: Número de ouvintes dos artistas do Last.fm

A Figura 4.5 representa gráficos da distribuição acumulada da popularidade dos artistas

escutados pelos ouvintes. Podemos observar que os valores da popularidade dos artistas com novidade (Figura 4.5(a)) são em geral um pouco menores que os valores para os artistas conhecidos (Figura 4.5(b)).



(a) Artistas com novidade

(b) Artistas conhecidos

Figura 4.5: Frequência acumulada da popularidade dos artistas escutados pelos ouvintes (já com valor transformado pelo log na base 10).

4.4.5 Preferências

Para mensurar o quanto o ouvinte gostou da novidade, foram utilizados duas métricas: a atenção total e o período de atenção. Como novidades podem ser descobertas em todo o Período de Experimento, alguns destes artistas possuem uma janela de tempo no experimento menor (artistas escutadas no final do Período de Experimento). Para contornar esse problema, utilizamos duas soluções. Primeiro, utilizamos como denominador no cálculo das métricas o número de semanas da Janela de Tempo de exposição à novidade, que vai da primeira semana que foi escutada a novidade até o fim do Período de Experimento. Segundo, como mencionado na Seção 4.2.1, apenas as novidades descobertas no Período de Observação foram consideradas na análise, mas todo o Período de Experimento foi utilizado para cálculo das métricas. Isso dá a cada novidade um mínimo de 6 meses de coleta de dados, que limita um possível viés para Janelas de Tempo pequenas.

A atenção total representa a atenção que o ouvinte deu para o artista no Período de Ex-

perimento. A atenção total do ouvinte para o artista é representada pelo total de número de execuções de músicas do artista que ele escutou no Período de Experimento, dividido pelo número de semanas de sua Janela de Tempo.

Já o período de atenção o tempo que o ouvinte deu atenção ao artista. Assim, é o número de semanas que o ouvinte escutou o artista dividido pelo número de semanas de sua Janela de Tempo.

Capítulo 5

Preferências dos ouvintes para diferentes aspectos de novidades

Após a coleta e filtragem dos dados, e da modelagem dos conceitos, partimos para responder as perguntas de pesquisa. Este capítulo abrange as duas primeiras perguntas, abordando as preferências dos ouvintes pelos aspectos - familiaridade e popularidade - da novidade, de forma geral e de forma individual.

5.1 Preferências gerais

Este trabalho visa entender como diferentes aspectos de novidades podem influenciar as preferências de ouvintes por elas. Para esta pesquisa, a primeira pergunta levantada foi: *Há uma correlação geral entre algum aspecto da novidade e as preferências dos ouvintes?* Para responder essa pergunta, calculamos a correlação entre cada aspecto da novidade - familiaridade e popularidade - e cada métrica de preferência - atenção total e período de atenção, de todas as novidades de todos os ouvintes juntas.

Com todas métricas das novidades em mãos, foi utilizado o método de correlação não-paramétrico de Spearman. O resultado gerado por este método pode variar de -1 a 1. Quanto mais próximo de 1, mais as variáveis estão correlacionadas positivamente - se uma cresce/decresce a outra cresce/decresce. Quanto mais próximo de -1, mais as variáveis estão correlacionadas negativamente - se uma cresce a outra decresce, e vice-versa. Se o valor estiver próximo a 0 não há correlação.

Dimensões	F	P
Período de atenção (PdA)	0,08	0,07
Familiarity (F)	-	0,08

Tabela 5.1: Correlação (Coeficiente de Spearman) entre aspectos da novidade e preferências, analisando todas as novidades juntas

A tabela 5.1 mostra o coeficiente de Spearman para cada par de aspecto da novidade / preferência. Como pode-se ver, todos os valores encontrados da correlação são próximos de zero. **Podemos concluir que no geral não existe uma correlação entre aspectos e preferências pela novidade, para todos os ouvintes juntos**. Por exemplo,não existe uma tendência clara pela qual os ouvintes gostam mais de uma novidade quando ela é mais familiar ou não-familiar. Nós levantamos duas hipóteses para explicar esse resultado:

1. Diferentes ouvintes possuem diferentes preferências para os aspectos das novidades. Nesta hipótese, cogitamos que diferentes ouvintes possuem diferentes preferências musicais. Assim, existe um grupo significativo de ouvintes que preferem novidades populares, outro grupo que preferem novidades não populares, etc. Colocando todos estes ouvintes juntos, a correlação geral vai ser próxima a zero.

2. Individualmente, os ouvintes gostam dos aspectos das novidades da mesma maneira. Cada ouvinte pode gostar, por exemplo, tanto de novidades familiares quanto não familiares, fazendo com que a correlação entre a preferência e o aspecto das novidades seja próxima a zero.

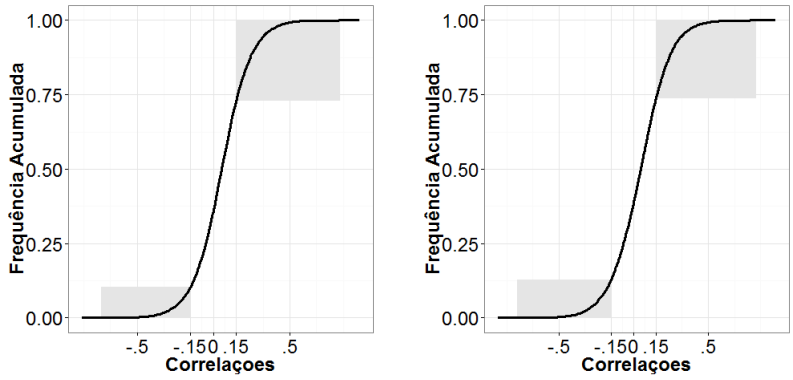
A seguir exploraremos estas hipóteses para saber qual se adequa a nossa situação.

5.2 Preferências individuais

O resultado da primeira pergunta de pesquisa e as hipóteses propostas nos levam à segunda pergunta de pesquisa: individualmente, os ouvintes preferem algum aspecto de novidade? Para saber se os ouvintes possuem alguma correlação entre os aspectos e as preferências das novidades, calculamos as correlações para cada sujeito individualmente. Cerca de 74%

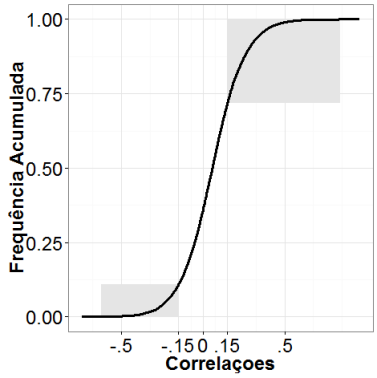
dos sujeitos possuem alguma correlação com valor maior que 0,15 ou menor que -0,15, e cerca de 26% possuem alguma correlação maior que 0,3 ou menor que -0,3. Desta maneira, individualmente, boa parte dos ouvintes possuem alguma correlação entre algum aspecto e alguma preferência da novidade. A Figura 5.1 mostra a distribuição acumulada dos valores das correlações para cada par aspecto / preferência.

Analisando estes dois resultados juntos, pode-se concluir que, apesar de não existir uma tendência geral quanto às preferências dos sujeitos para os diferentes aspectos das novidades, **a maior parte dos sujeitos possuem alguma preferência para algum aspecto de novidade no seu comportamento**. Isso sugere a presença de diferentes tipos de ouvintes nos nossos dados. Para identificar estes tipos, foi utilizado um algoritmo de agrupamento nos sujeitos, que será discutido no próximo capítulo.

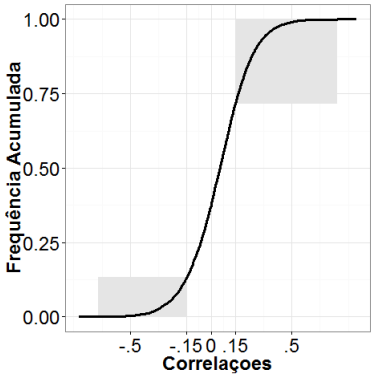


(a) Atenção total e familiaridade

(b) Atenção total e popularidade



(c) Período de atenção e familiaridade



(d) Período de atenção e popularidade

Figura 5.1: Distribuição do coeficiente Spearman de correlação entre os aspectos das novidades e as preferências de cada ouvinte dos dados. As áreas sombreadas representam a parte dos sujeitos com correlação maior que 0.15 e menor que -0.15.

Capítulo 6

Grupos de usuários para diferentes aspectos de novidade

De acordo com os resultados do Capítulo 5, há a evidência de diferentes tipos de sujeitos nos nossos dados, de acordo com as preferências pelos aspectos das novidades. Isto nos leva a terceira pergunta de pesquisa: Existem grupos de ouvintes relacionados com as preferências pelos aspectos das novidades baseadas no perfil? Para identificar estes grupos, foi realizado um agrupamento nos dados que caracterizam os sujeitos. Esta análise será descrita neste capítulo.

6.1 Conjunto de sujeitos

Primeiramente, a análise foi realizada com os dados dos mesmos sujeitos utilizados na análise das preferências dos ouvintes para os aspectos das novidades (Capítulo 5). Depois, para estudar especificamente os sujeitos com alguma correlação entre preferência e aspectos da novidade, foi feita a análise com dois subconjuntos: o primeiro subconjunto consiste em sujeitos com valor de correlação entre algum aspecto e preferência da novidade maior que 0,15 ou menor que -0,15; o segundo consiste em sujeitos com valor de correlação maior que 0,3 ou menor que -0,3. Os resultados encontrados nas três análises foram similares. Portanto, mostraremos apenas os resultados da primeira análise.

6.2 Dados que caracterizam sujeitos

O objetivo da análise atual é identificar os grupos de ouvintes baseados nas preferências pelos aspectos das novidades comparadas com seus hábitos musicais. A ideia de investigar as métricas relacionadas às novidades e aos hábitos musicais é para saber se o comportamento do ouvinte por novidade é semelhante ao seu comportamento no geral. Foram escolhidas 5 métricas de caracterização dos sujeitos, que podem ser divididas em dois grupos: métricas relacionadas com novidades e métricas relacionadas com os hábitos musicais.

1. **Relacionadas com novidades**: Métricas que caracterizam os sujeitos a partir de suas preferências pelos aspectos das novidades.

    - *Correlação entre a familiaridade das novidades e a atenção total dedicada a elas durante o experimento.*
    - *Correlação entre popularidade das novidades e a atenção total dedicada a elas durante o experimento.*

    **Calculamos a correlação (coeficiente de Spearman) entre a atenção total e o período de atenção, e o valor encontrado foi de 0,71. Como ambas as variáveis estão correlacionadas, decidimos utilizar no algoritmo de agrupamento apenas as correlações que envolvem a atenção total.*

2. **Relacionadas com os hábitos musicais**: Métricas que caracterizam os hábitos musicais do sujeito e que podem ser contrastadas com as métricas das novidades.

    - *Ecleticidade* A ecleticidade representa o quão diferente os artistas do perfil do ouvinte são, de acordo com seus descritores. Sua definição e cálculo foi feito na Subseção 4.4.2. A ecleticidade está relacionada com a familiaridade da novidade, pois quanto mais eclético um sujeito for, maior a probabilidade dele ser familiar a vários tipos de novidades.
    - *Popularidade média dos artistas do perfil do ouvinte* Esta métrica é útil para comparar com a popularidade das novidades escutadas pelo ouvinte.

- *Proporção de novidades escutadas pelo ouvinte no Período de Observação* Com esta métrica pode-se identificar se o ouvinte possuiu o hábito de escutar muitas ou poucas novidades, no Período de Observação.

6.3 Algoritmo de agrupamento

Para fazer o agrupamento dos sujeitos foi utilizado o algoritmo de agrupamento aglomerativo hierárquico. Como descrito na Seção 3.3.2, este tipo de algoritmo necessita de uma métrica de dissimilaridade entre os pares de sujeitos e um critério de união que especifica quais grupos unir em cada passo.

Para calcular a dissimilaridade entre os sujeitos, primeiro foram calculadas as métricas descritas na Seção 6.2. Após o cálculo, estes dados foram normalizados, baseados no Z-Score. Então, a dissimilaridade entre dois sujeitos foi calculada a partir da distância euclidiana, onde cada sujeito é representado por um vetor contendo as 5 métricas normalizadas. Formalmente, sejam S o conjunto de vetores com dados normalizados que representam cada sujeito; e $x \in S$ e $y \in S$ vetores de dois sujeitos de S. x_i representa a posição i do vetor x, no caso uma das 5 métricas que caracterizam os sujeitos. A dissimilaridade entre os sujeitos, representada pela dissimilaridade $\text{disSuj}(x, y)$ dos vetores que os representam é dado pela Equação 6.1. Como critério de união, foi utilizado o método Ward [27].

$$\text{disSuj}(x, y) = \sqrt{\sum_{i=1}^{5}(x_i - y_i)^2} \tag{6.1}$$

6.4 Escolha do número de grupos

O método de agrupamento hierárquico não expõe explicitamente o número de grupos resultantes. A cada passo, o algoritmo une dois grupos, até que todos os sujeitos estejam em um só grupo. Uma abordagem para definir a melhor configuração de número de grupos é o método do joelho [25], ao plotar um gráfico onde o eixo X é o número de grupos e o Y um critério de avaliação. A Figura 6.1 mostra a distância média dentro dos grupos para cada configuração de número de grupos. O método de joelho determina escolher uma configuração de grupos que não adicionem muita heterogeneidade, evidenciado a partir da curvatura

máxima do gráfico (joelho). Pela figura, a distância média dentro dos grupos começa a aumentar vertiginosamente nas configurações com menos de 7 grupos. Assim, analisamos as configuração com 7 e 8 grupos.

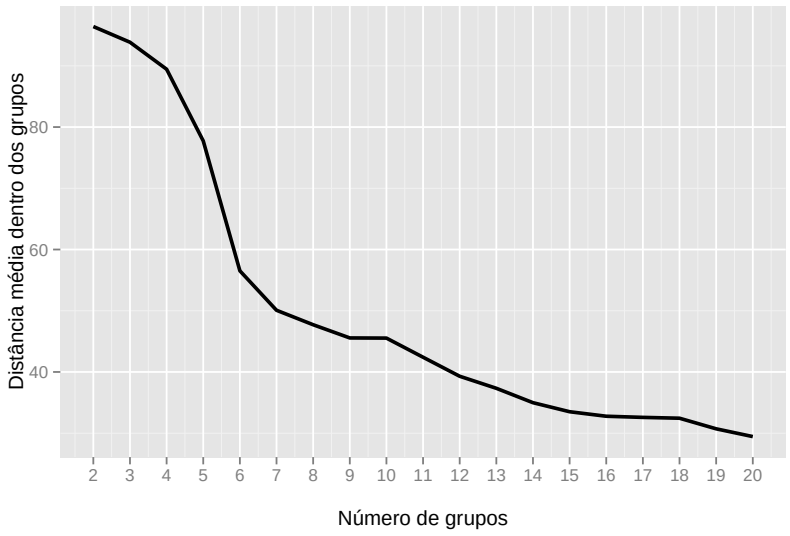


Figura 6.1: Número de grupo X Distância média dentro dos grupos. O joelho do gráfico está em torno da configuração com 7 grupos.

A Tabela 6.1 mostra os valores dos centróides para a solução de 7 grupos e o centróide criado na solução de 8 grupos (Centróide 8). Como podemos ver, o Centróide 2 é similar ao Centróide 8. Desta maneira, escolhemos a configuração com 7 grupos.

6.5 Grupos

Após a escolha de 7 grupos de ouvintes, utilizamos o centróide de cada grupo para analisar as suas principais características. A Figura 6.2 representa os valores dos centróides normalizados. Podemos dividir os grupos em dois tipos: o primeiro, onde as características que se destacam são as relacionadas com as preferências pelos aspectos das novidades e o segundo, onde as características que se destacam são as relacionadas com os hábitos musicais dos ouvintes. De acordo com cada centróide, rotulamos os grupos da seguinte forma:

1. Grupos marcados pelas preferências pelos aspectos das novidades

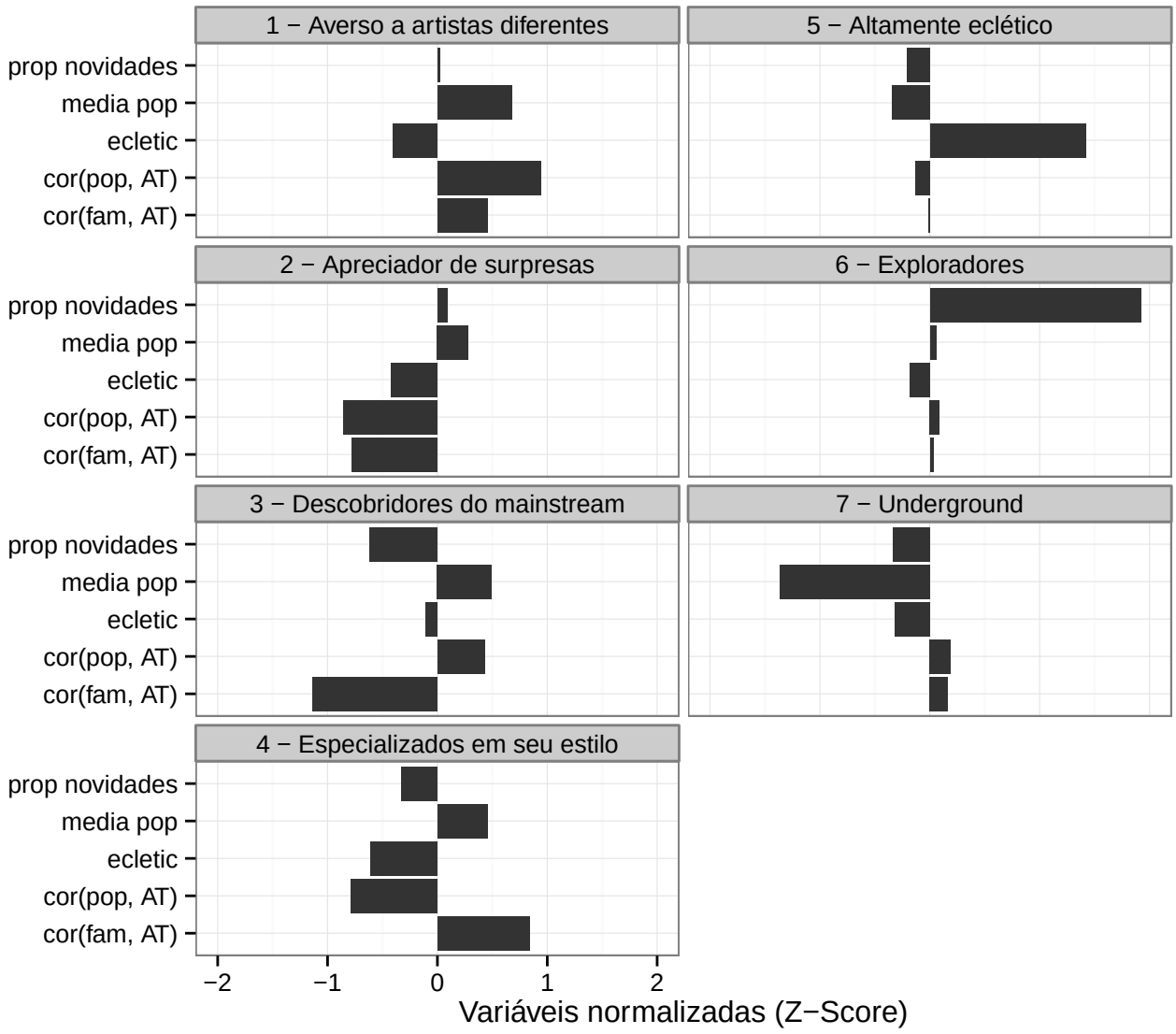


Figura 6.2: Centróides dos 7 grupos encontrados na análise. As métricas estão normalizadas pelo z-score, onde zero representa a média de todos os ouvintes, e a unidade de variação é um desvio padrão, para cada métrica. No eixo vertical, *fam* significa a familiaridade, *pop* significa popularidade, *AT* significa atenção total.

Centro	cor(fam,TA)	cor(pop,TA)	ecletic	media pop	prop novid
Centro 1	0.46	0.94	-0.4	0.67	0.02
Centro 2*	-0.01	-0.13	1.42	-0.34	-0.21
Centro 3	0.03	0.09	-0.18	0.06	1.92
Centro 4	0.16	0.19	-0.32	-1.36	-0.33
Centro 5	-0.77	-0.85	-0.42	0.28	0.09
Centro 6	-1.13	0.43	-0.11	0.49	-0.62
Centro 7	0.84	-0.79	-0.61	0.46	-0.33
Centro 8**	0.23	-0.49	1.24	0.03	-0.74

Tabela 6.1: Centróides para as configurações com 7 grupos e com 8 grupos

(a) Averso a artistas diferentes (ou acomodado) [total de ouvintes: 2317 (20%)]: Maior grupo com característica marcante pelas preferências pelos aspectos das novidades, formado por ouvintes que preferem novidades familiares e populares, além de possuírem hábitos musicais marcados por artistas populares.

(b) Apreciador de surpresas [total de ouvintes: 1859 (17%)]: Ouvintes preferem novidades não-familiares e pouco populares.

(c) Descobridores do mainstream [total de ouvintes: 1022 (8%)]: Ouvintes que preferem novidades não-familiares e populares.

(d) Especializados em seu estilo [total de ouvintes: 1467 (14%)]: Ouvintes que preferem novidades pouco populares e familiares, além de possuirem pouca ecleticidade.

2. Grupos marcados pelas características dos hábitos musicais

(a) Altamente eclético [total de ouvintes: 2456 (21%)]: Maior grupo de todos, formado por ouvintes que possuem alta ecleticidade

(b) Exploradores [total de ouvintes: 1047 (9%)]: Ouvintes que possuem alta proporção de novidades escutadas durante o Período de Observação

(c) Underground [total de ouvintes: 1281 (11%)]: Ouvintes com hábito musical marcado por artistas pouco populares.

6.6 Discussão dos grupos encontrados

Encontramos 4 grupos que possuem como características marcantes as preferências pelos aspectos das novidades. Coincidentemente, foram encontrados grupos com todas as combinações possíveis de preferências pelos aspectos.

O maior destes 4 grupos é o que chamamos *Averso a artistas diferentes*. É um grupo de ouvintes que preferem novidades familiares e populares, além de possuírem hábitos musicais marcados por artistas populares, pouca ecleticidade e proporção mediana de novidades escutadas. É um tipo de ouvinte que não procura expandir seu perfil musical, preferindo escutar o que está na mídia do que ele habitualmente já escuta. A Figura 6.3 representa o perfil de um ouvinte *Averso a artistas diferentes*. Nota-se que as novidades mais preferidas são as mais familiares e mais populares.

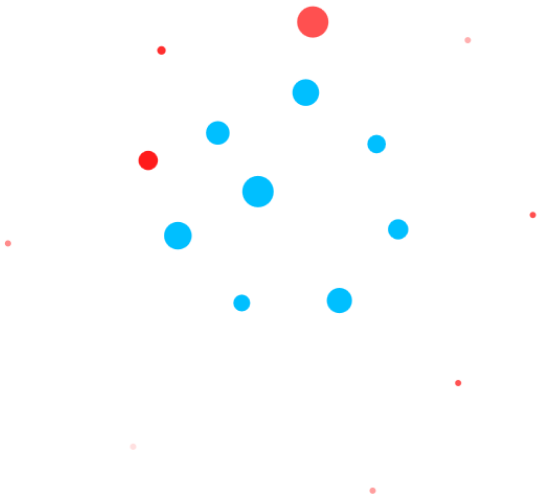


Figura 6.3: Perfil de um ouvinte *Averso a artistas diferentes*. Os círculos azuis são os clusters e os vermelhos as novidades. A opacidade dos círculos vermelhos está relacionada com a preferência do ouvinte pela novidade em questão. As novidades mais preferidas são as mais familiares e populares.

Os outros 3 grupos dos marcados pelas preferências pelos aspectos das novidades possuem uma distribuição mais homogênea do número de ouvintes. Opostos aos *Aversos a artistas diferentes*, os *Apreciadores de surpresas* preferem novidades não familiares e não populares, além de possuir pouca ecleticidade. Desta maneira, estes ouvintes normalmente escutam artistas bem parecidos, mas tentam aumentar esse leque de artistas do perfil prefe-

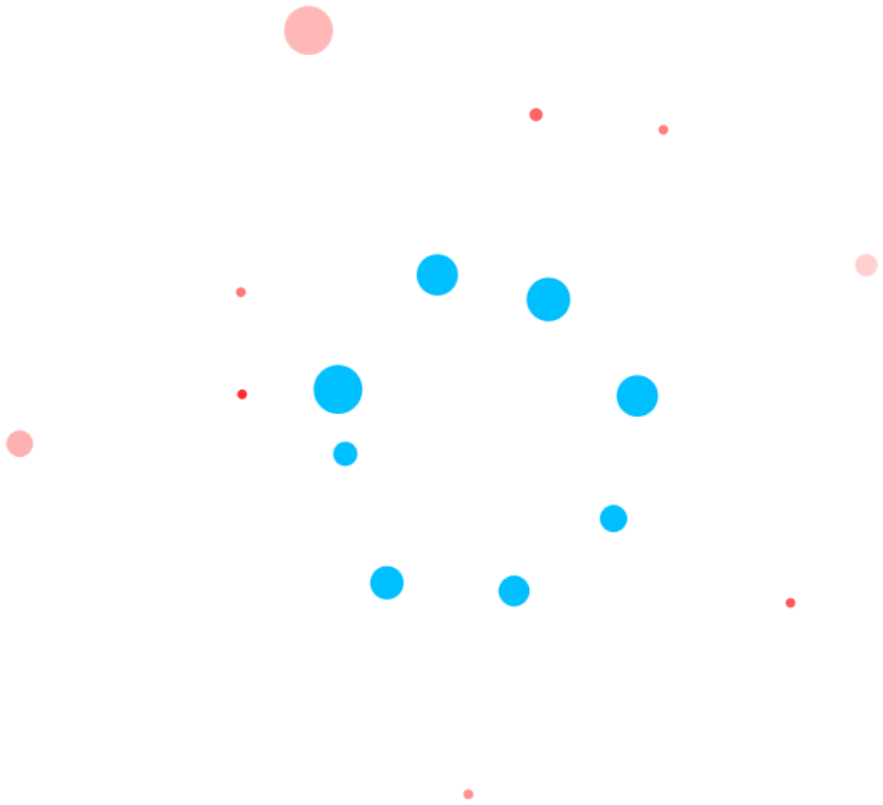


Figura 6.4: Perfil de um ouvinte *Apreciador de surpresas*. As novidades mais preferidas são as menos familiares e menos populares.

rindo novidades não familiares e não populares. Eles preferem surpresas, artistas diferentes do que já escutaram. A Figura 6.4 representa o perfil de um ouvinte *Apreciador de surpresas*. Nota-se que as novidades mais preferidas são as menos familiares e menos populares.

Os *Descobridores do mainstream* preferem novidades populares, que estão na mídia, mesmo não sendo familiares. Uma possível explicação seria ouvintes que escutam o que está nas paradas das rádios, sem se importar se são parecidos com o que ele escutava antes ou não. Já os *Especializados em seu estilo* preferem novidades familiares e pouco populares, além de possuírem pouca ecleticidade. Esse tipo de ouvinte são fechados no seu nicho de estilos musicais, e ou não preferem o que está na mídia destes estilos, ou já escutaram tudo que está na mídia e agora querem expandir para artistas não populares destes estilos.

Analisando os 3 grupos restantes, o *Altamente eclético* é o grupo de ouvintes com maior ecleticidade, comparando com os demais. Nota-se que neste grupo existem diferentes sujeitos, por que as outras métricas são próximas a zero. O grupo de *Exploradores*, formado por ouvintes com alta proporção de novidades escutadas no período de observação, também possui essa característica do *Altamente eclético*. Por fim, o grupo *Underground* é formado por ouvintes que têm hábito de escutar artistas pouco populares no geral.

Capítulo 7

Comparação das novidades com os artistas conhecidos

Até este momento trabalhamos apenas com novidades. Porém, será que as preferências dos ouvintes pelas novidades são as mesmas que pelos artistas conhecidos? Caso esta hipótese seja verdade, o comportamento dos ouvintes para estes dois âmbitos, novidades e não-novidades, seriam similares, podendo estender os resultados das novidades para os itens conhecidos e vice-versa. Para responder esta pergunta expandimos nossos experimentos para englobar também os artistas conhecidos, tentando responder a seguinte pergunta de pesquisa: *As relações entre as preferências dos ouvintes e os aspectos das novidades são as mesmas que as relações entre as preferências dos ouvintes e os aspectos dos artistas já conhecidos?*

7.1 Seleção de sujeitos

Para esta análise, utilizamos os mesmos 11.989 sujeitos das análises anteriores e aplicamos um filtro, para selecionar os sujeitos que foram expostos a um número mínimo de artistas conhecidos que permitisse a investigação de relações entre as características destes artistas e as preferências. Filtro similar foi descrito na Seção 4.2.3 para a seleção de sujeitos que foram expostos a um número mínimo de novidades.

Desta maneira, foram excluídos os usuários que escutaram menos de 10 artistas conhecidos no Período de Observação. Após esta exclusão ficamos com uma amostra de 10.207 sujeitos para os experimentos.

7.2 Características dos artistas conhecidos

Como dito na Seção 4.2.1, os artistas conhecidos são os artistas escutados no Período de Observação que já foram escutados previamente no Histórico Inicial do ouvinte. Para eles foram calculadas 3 das 4 métricas que foram calculadas para os artistas com novidade (Capítulo **??**): familiaridade e popularidade (aspectos) e total de atenção (preferência).

Para calcular as relações entre os aspectos dos artistas conhecidos e as preferências pelos ouvintes, utilizamos a mesma metodologia descrita no Capítulo 5. Utilizamos o método de correlação não-paramétrico de Spearman entre cada par de aspecto / preferência.

7.3 Comparação entre relações das preferências e aspectos das novidades e dos artistas conhecidos

Para guiar nossos experimentos, decidimos especificar a quarta pergunta de pesquisa, que é mais genérica. Assim, tentamos responder duas perguntas de pesquisa.

A primeira *As relações entre as preferências dos ouvintes e os aspectos das novidades são significantemente diferentes das relações entre as preferências dos ouvintes e os aspectos dos artistas já conhecidas?* Se as relações comparadas forem significantemente iguais, poderíamos estender os resultados das novidades para o comportamento geral do ouvinte. Se forem diferentes, corroboraremos a importância da análise separada de itens com novidades e itens conhecidos.

Para cada um dos 10.207 sujeitos foram calculados os valores das correlações entre a familiariade e o total de atenção e entre a popularidade e o total de atenção, tanto para os artistas com novidade quanto para os artistas conhecidos. Os detalhes destes cálculos foram apresentados no Capítulo 5 e na Seção 7.2. Desta maneira cada ouvinte possui pares de correlações (Tabela 7.1), onde um valor do par é referente aos artistas com novidade e o outro os artistas conhecidos. Por exemplo, cada ouvinte possui um valor de correlação entre a familiaridade e o total de atenção para as novidades e um valor de correlação entre a familiaridade e o total de atenção para os artistas conhecidos.

Desta maneira, para responder a pergunta exposta no parágrafo anterior, utilizamos um teste-T pareado, para cada par de correlação novidade / artista conhecido correspondente de

todos os ouvintes. O intuito desta análise é descobrir se, por exemplo, o valor da correlação entre a familiaridade e o total de atenção para as novidades e para os artistas conhecidos são significantemente diferentes.

Artistas com novidade	Artistas conhecidos
cor(familiaridade, atenção total)	cor(familiaridade, atenção total)
cor(popularidade, atenção total)	cor(popularidade, atenção total)

Tabela 7.1: Correlações calculadas para artistas com novidade e artistas conhecidos.

A Tabela 7.2 mostra o p-valor e o intervalo de confiança da média das diferenças para cada par de aspecto / preferência. Em ambas métricas comparadas, a probabilidade da média das diferenças entre os valores para os artistas com novidade e os artistas conhecidos ser igual a zero é baixíssima (p-valor < 0.0001). Desta maneira, podemos afirmar que as correlações no geral são diferentes. Ou seja, as relações dos ouvintes entre as preferências e os aspectos, para artistas com novidade e artistas conhecidos, são diferentes.

Métricas	P-Valor	Intervalo de Confiança (α = 0.05)
Familiaridade / Período de atenção	< 0.0001	[-0,14 -0,13]
Popularidade / Período de atenção	< 0.0001	[-0,04 -0,03]

Tabela 7.2: Teste-T pareado entre as correlações dos aspectos e preferências dos ouvintes para artistas com novidade e artistas conhecidos.

Outro fato que podemos observar da tabela é que o intervalo de confiança, com 95% de confiança, para as diferenças das correlações entre familiaridade e atenção total para artistas com novidade e artistas conhecidos é [-0,14 -0,13], sugerindo que as preferências pelos artistas conhecidos familiares são maiores que as preferências pelos artistas com novidade familiares. Uma possível explicação seria que os artistas conhecidos mais familiares são os artistas que o ouvinte mais escutou no seu histórico. Assim, naturalmente a preferência por eles é bem maior.

Já o intervalo de confiança, com 95% de confiança, para as diferenças das correlações entre popularidade e atenção total para artistas com novidade e artistas conhecidos é [-0,04 -0,03]. Também sugere que as preferências pelos artistas mais populares conhecidos são

maiores que pelos artistas mais populares com novidade, apesar de que este intervalo de confiança está mais próximo de zero que o intervalo de confiança anterior.

No geral, podemos concluir que as preferências dos ouvintes por artistas mais familiares e mais populares é menor se os artistas forem novidades do que se eles forem artistas conhecidos. É como se a preferência por artistas conhecidos já esteja consolidada e que ao escutar novidades os ouvintes tendem a explorar mais, escutando no mesmo período de tempo mais artistas menos familiares e menos populares que os artistas conhecidos.

Já a segunda pergunta de pesquisa, *Existe correlação entre as relações das preferências pelos aspectos dos artistas com novidade e as relações das preferências pelos aspectos dos artistas conhecidos?* Próximo passo foi verificar se existiam correlações entre as relações calculadas para os artistas com novidade e para os artistas conhecidos. Será que quanto maior a correlação de um ouvinte entre familiaridade e atenção total para artistas com novidade, por exemplo, maior a correlação entre a familiaridade e atenção total para artistas conhecidos? Para responder a pergunta de pesquisa, utilizamos o método de correlação não-paramétrico de Spearman.

Novidades X Artistas conhecidos	cor(fam., aten. tot.)	cor(pop. , aten. tot.)
cor(familiaridade , atenção total)	0,05	-0.02
cor(popularidade , atenção total)	-0.04	0,09

Tabela 7.3: Correlação (Coeficiente de Spearman) entre correlações calculadas para novidades (linhas) e artistas conhecidos (colunas)

Como podemos observar na Tabela 7.3, a correlação é baixa em todos os casos. Não podemos afirmar, por exemplo, que dado dois ouvintes, se um possuir maior correlação entre familiaridade e período de atenção para novidades que o outro ouvinte ele terá maior chance de ter maior correlação entre familiaridade e período de atenção para artistas conhecidos, e vice-versa.

Esta falta de correlação geral pode ser ocasionada pela diversidade de comportamentos para novidades e artistas conhecidos dos ouvintes. Levantamos a hipótese de que, por exemplo, existem ouvintes que possuem uma preferência mais forte por novidades familiares que por artistas conhecidos familiares, enquanto outros não. Hipótese semelhante a esta

foi exposta no Capítulo 5, que nos levou a criar perfis de ouvintes baseados nas relações entre preferências e os aspectos das novidades (Capitulo 6). Desta maneira, fizemos uma investigação parecida para testar a hipótese.

7.4 Grupos de ouvintes baseados na diferença das relações entre preferências e aspectos das novidades e dos artistas conhecidos

Como as correlações encontradas na segunda pergunta de pesquisa da seção anterior foram baixas, seguindo a metodologia dos Capítulos 5 / 6 tentamos encontrar perfis de ouvintes, com o intuito de responder a seguinte pergunta: *Quais os diferentes perfis de ouvintes baseados nas diferenças das relações preferências / aspectos das novidades e dos itens conhecidos?*

Primeiro passo foi calcular duas diferenças, para cada ouvinte: primeiro, a diferença entre as correlações da familiaridade e atenção total das novidades e dos artistas conhecidos (linha 1 da Tabela 7.1) e segundo, a diferença entre as correlações da popularidade e atenção total das novidades e dos artistas conhecidos (linha 2 da Tabela 7.1) .

Com estas diferenças em mãos utilizamos o algoritmo de agrupamento aglomerativo hierárquico. A escolha do número de grupos foi baseada no método do joelho. Na imagem 7.1 podemos observar um início de joelho entre as configurações com 4 e 6 grupos. Comparando os centróides das configurações de 6 e 7 grupos (Tabela 7.4), podemos notar que o centróide adicionado na configuração de 7 grupos (Centróide 7) é similar ao Centróide 5. Assim, como adicionar um grupo na configuração 6 não muda muito, escolhemos a configuração de 6 grupos.

Após a escolha dos 6 grupos de ouvintes, utilizamos o centróide de cada grupo para analisar as suas principais características. A Figura 7.2 representa os valores dos centróides.

Podemos dividir os grupos em quatro tipos:

1. Tipo I - Preferências similares para novidades e itens conhecidos [total de ouvintes: 2.427 (23%)]: Composto pelo grupo 1, o centroide deste evidencia que as diferenças entre as correlações, dos dois pares de preferência / aspecto, calculadas para novidades

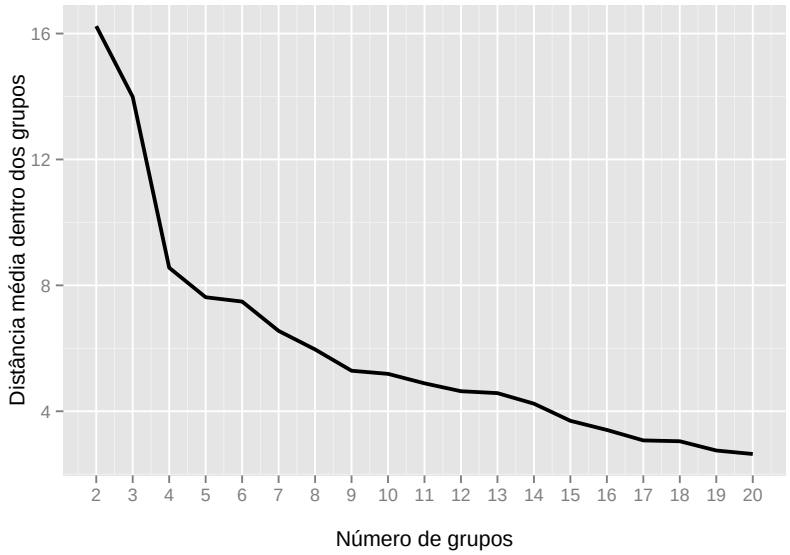


Figura 7.1: Número de grupo X Distância média dentro dos grupos.

Centro	cor(AT,fam)	cor(AT,pop)
Centro 1	-0.02	0.08
Centro 2	0.06	0.35
Centro 3	-0.35	-0.29
Centro 4	0.05	-0.19
Centro 5*	-0.31	0.06
Centro 6	-0.15	-0.15
Centro 7**	-0.26	0.07

Tabela 7.4: Centróides para a configuração com 6 grupos e com 7 grupos

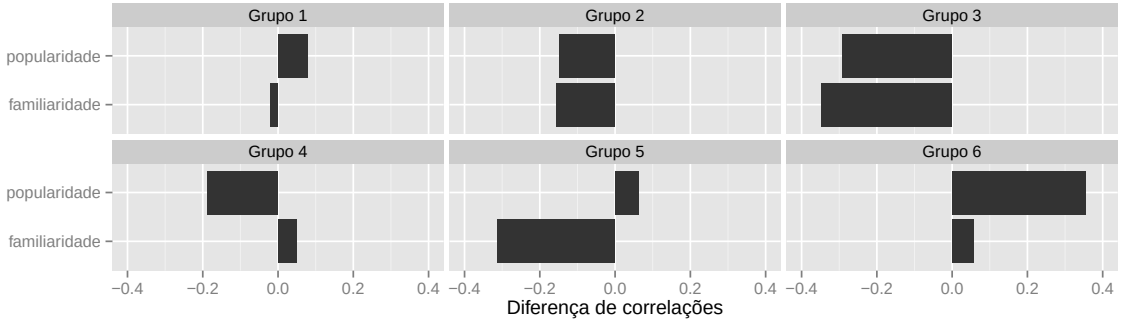


Figura 7.2: Centróides dos 6 grupos encontrados na análise da diferença das correlações

e itens conhecidos, são próximas a zero.

2. Tipo II - Correlações para itens conhecidos maiores que para novidades [total de ouvintes: 2.833 (27.7 %)]: O segundo tipo é composto pelos grupos 2 e 3, onde as correlações, dos dois pares de preferência / aspecto, são maiores para os itens conhecidos que para as novidades, seja em maior (grupo 2) ou menor (grupo 3) grau.

3. Tipo III - Correlação para itens conhecidos maior que para novidades em apenas um aspecto [total de ouvintes: 4.529 (44.3 %)]: composto pelos grupos 4 e 5, apenas um par preferência / aspecto possui correlação para itens conhecidos maior que para novidades, enquanto que o outro par é próximo a zero. O grupo 4 é formado no geral por ouvintes que possuem correlações entre preferência e familiaridade para itens conhecidos maiores que para novidades. Já o grupo 5 é formado no geral por ouvintes que possuem correlações entre preferência e popularidade para itens conhecidos maiores que para novidades.

4. Tipo IV - Correlação das preferências e popularidade para novidades maior que para itens conhecidos [total de ouvintes: 418 (5%)]: por último, este tipo de ouvinte possui correlação entre preferência e popularidade maior para novidades que para artistas conhecidos. Composto pelo grupo 6.

   Boa parte dos grupos (4 dos 6) são formados no geral por ouvintes que possuem pelo menos correlação de algum par preferência / aspecto maior para itens conhecidos que para novidades. Isso corrobora os resultados encontrados na Seção 7.3, que mostra que no geral os ouvintes possuem correlações maiores para itens conhecidos que para novidades. Cerca de 23% dos ouvintes estão no grupo de comportamento parecido para novidades e itens conhecidos. Assim, poderíamos tentar prever comportamento para novidades baseados no comportamento de itens conhecidos para apenas 1/4 dos ouvintes, enquanto os demais precisariam de um tratamento diferenciado para estes dois tipos de comportamentos.

Capítulo 8

Conclusão

Neste capítulo apresentamos um resumo geral da nossa pesquisa sobre os aspectos e preferências das novidades musicais. A partir deste resumo, discutimos os resultados obtidos e as implicações.

8.1 Resumo

O principal objetivo da pesquisa foi analisar as novidades de forma multidimensional, de acordo com dois aspectos - a popularidade e a familiaridade. A análise foi feita relacionando estes aspectos com as preferências dos ouvintes pelas novidades.

Na primeira parte da pesquisa analisamos as correlações entre as preferências dos ouvintes e os aspectos das novidades, para as novidades de todos os sujeitos juntos. Foi descoberto que não há uma correlação entre as preferências e os aspectos, analisando todos os ouvintes juntos. Este resultado motivou a análise das correlações para cada indivíduo. Esta segunda análise mostrou que individualmente os ouvintes possuem preferências relacionadas com os aspectos das novidades. Assim, fizemos uma análise para encontrar diferentes grupos de ouvintes, baseados nas relações entre preferências dos ouvintes e aspectos das novidades.

Já a segunda parte da pesquisa envolveu comparações das relações de preferências dos ouvintes e aspectos das novidades com as relações para os artistas conhecidos. Descobrimos que estas relações são diferentes para os artistas com novidade e os artistas conhecidos. As relações que envolvem a familiaridade e a popularidade são, em média, maiores para os artistas conhecidos que para as novidades. Além disso, fizemos uma análise para encontrar

grupos de ouvintes baseados nas diferenças entre as relações para as novidades e artistas conhecidos.

8.2 Implicações

O resultado geral da pesquisa foi dar suporte ao tratamento de novidades de maneira multidimensional. Os resultados em geral mostraram que os ouvintes possuem preferências diferentes para diferentes aspectos das novidades. Esse resultado apoia os modelos propostos por Vargas et. al [26] e Beloggin et. al [3].

Esta visão de que diferentes ouvintes possuem preferências diferentes para os aspectos das novidades implica na importância de soluções no âmbito de novidades musicais que levem em consideração estas diferenças. Por exemplo, construtores de sistemas de recomendação de novidades musicais poderiam aperfeiçoar os algoritmos para levar em conta as preferências do ouvinte tanto pela familiaridade quanto pela popularidade das novidades.

Outra solução, principalmente motivada pela presença dos grupos de ouvintes, envolve o design de sites e/ou ferramentas de música para que seja um design específico para cada grupo de ouvinte. Outras características, como redirecionamento de notícias e recomendação de shows/festivais, poderiam ser incrementadas neste sistema levando em conta estes aspectos das novidades.

Também utilizando esta evidência de diferentes grupos, sistemas poderiam incorporar as preferências pelos aspectos das novidades em recomendação de vizinhos. Vizinhos musicais são ouvintes que possuem gostos similares. Desta maneira, a recomendação de vizinhos é um tipo especial de recomendação, onde em vez de itens, são recomendados ouvintes que compartilham as mesmas preferências. Neste caso, seriam compartilhados ouvintes com preferências similares pelos aspectos das novidades.

Por fim, foi mostrado que, de maneira geral, os ouvintes possuem diferentes relações entre as preferências e os aspectos das novidades e as preferências e os aspectos dos artistas conhecidos. Isso implica que os ouvintes podem possuir diferentes comportamentos frente a artistas conhecidos e artistas com novidade, sugerindo um tratamento específico para cada um destes âmbitos.

Bibliografia

[1] Apple unveils new itunes. https://www.apple.com/pr/library/2012/09/12Apple-Unveils-New-iTunes.html, 2012.

[2] Leena Arhippainen and Seamus Hickey. Classifying music user groups and identifying needs for mobile virtual music services. page 191, 2011.

[3] Alejandro Bellogín, Iván Cantador, and Pablo Castells. A study of heterogeneity in recommendations for a social music service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '10, pages 1–8, New York, NY, USA, 2010. ACM.

[4] Òscar Celma. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[5] Òscar Celma and Pedro Cano. From hits to niches?: Or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2Nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, NETFLIX 2008, pages 5:1–5:8, New York, NY, USA, 2008.

[6] Òscar Celma and Perfecto Herrera. A new approach to evaluating novel recommendations. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 179–186, New York, NY, USA, 2008. ACM.

[7] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173, February 2011.

[8] Fishbach, Ayelet, Rebecca K. Ratner, and Ying Zhang. Inherently loyal or easily bored? nonconscious activation of consistency versus variety-seeking behavior. *Journal of Consumer Psychology*, 21(1):38–48, 2011.

[9] Leo A. Goodman. Snowball sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, 03 1961.

[10] D. J. Hargreaves and a. C. North. The Functions of Music in Everyday Life: Redefining the Social in Music Psychology. *Psychology of Music*, 27(1):71–83, April 1999.

[11] David J. Hargreaves, Dorothy Miell, and Raymond. *WHAT ARE MUSICAL IDENTITIES, AND WHY ARE THEY IMPORTANT?* Oxford University Press, USA, 2002.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Hierarquical clustering. In *The elements of statistical learning: data mining, inference and prediction*. Springer, 2009.

[13] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004.

[14] J. Jacoby and D. B Kyner. Brand loyalty vs. repeat purchasing behavior. *Journal of Marketing Research*, 10:1–19, 1973.

[15] David Jennings. *Net, Blogs and Rock 'n' Roll: How Digital Discovery Works and What it Means for Consumers*. Nicholas Brealey Publishing, 2007.

[16] V. J. Konecni. Social interaction and musical preference. *The psychology of music*, pages 497–516, 1982.

[17] Paul Lamere and Elias Pampalk. Social tags and music information retrieval. In Juan Pablo Bello, Elaine Chew, and Douglas Turnbull, editors, *ISMIR*, page 24, 2008.

[18] Audrey Laplante. Social capital and music discovery: An examination of the ties through which late adolescents discover new music. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 341–346. University of Miami, 2011.

[19] S. R. Maddi. The pursuit of consistency and variety. In R. P. Abelson et al, editor, *Theories of cognitive consistency*. Rand McNally, Chicago, 1968.

[20] Sean A. Munson and Paul Resnick. Presenting diverse political opinions: How and how much. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1457–1466, New York, NY, USA, 2010. ACM.

[21] Makoto Nakatsuji, Yasuhiro Fujiwara, Akimichi Tanaka, Toshio Uchiyama, Ko Fujimura, and Toru Ishida. Classical music for rock fans?: novel recommendations for expanding user interests. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 949–958, New York, NY, USA, 2010. ACM.

[22] MichaelJ. Pazzani and Daniel Billsus. Content-based recommendation systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 325–341. Springer Berlin Heidelberg, 2007.

[23] Rebecca K. Ratner, Barbara E. Kahn, and Daniel Kahneman. Choosing Less Preferred Experiences For the Sake of Variety. *Journal of Consumer Research*, 26:1–15, 1999.

[24] Paul Resnick and Hal R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, March 1997.

[25] Robert Thorndike. Who belongs in the family? *Psychometrika*, 18:267–276, 1953.

[26] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, RecSys '11, pages 109–116, New York, NY, USA, 2011. ACM.

[27] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[28] Morgan Ward, Joseph Goodman, and Julie Irwin. The same old song: The power of familiarity in music choice. *Marketing Letters*, 25(1):1–11, 2014.

[29] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 13–22, New York, NY, USA, 2012. ACM.

[30] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.