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

Título de sua Dissertação

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: Linha de Pesquisa

Nazareno Ferreira de Andrade
(Orientador)

Campina Grande, Paraíba, Brasil

Resumo

Seu resumo aqui

Abstract

Abstract Here

Agradecimentos

Agradecimentos

Conteúdo

Lista de Símbolos

HI - *Histórico Inicial*

PO - *Período de Observação*

PE - *Período do Experimento*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A procura e descoberta de novas músicas e artistas é um aspecto importante no consumo musical. Maddi [2] argumenta que consumidores em geral possuem um *impulso interno*, que tem como finalidade descobrir novas experiências, afim de criar novos sentimentos e emoções.

E este consumo musical mudou nos últimos anos. Serviços de streaming como, Spotify, Youtube, Soundcloud, rádios online como a do Last.Fm, até mesmo os sites de compra de música digital, como Itunes, 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. Muitas vezes o ouvinte acaba não descobrindo novidades que seriam de seu interesse.

Para resolver este problema, tanto sistemas comerciais como a academia apresentam soluções, principalmente incorporando a sistemas de recomendação o domínio da *novidade*. Mas boa parte destas abordagens tratam novidade de forma unidimensional. Podemos caracterizar uma novidade com diferentes dimensões, ou aspectos, como a familiaridade e a popularidade. Por exemplo, um ouvinte pode preferir novidades similares, ou familiares, a músicas que ele costuma escutar, mas preferir novidades não populares, e vice-versa.

Com o intuito de expandir o entendimento sobre as novidades, conduzimos uma análise sobre o impacto dos diferentes aspectos das novidades para as preferências de ouvintes de música. Para auxiliar esta análise, procuramos responder as seguintes perguntas de pesquisa:

1. Há alguma relação geral entre algum aspecto das novidades e as preferências dos ou-

vintes?

2. Individualmente, os usuários preferem algum aspecto das novidades?

3. Existem grupos de ouvintes baseados nas preferências pelos aspectos das novidades comparadas com seus hábitos musicais?

4. 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 das músicas já conhecidas?

A primeira pergunta tenta 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 outras músicas escutadas anteriormente (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". A segunda pergunta é uma especificação da primeira, em um âmbito individual. 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."

A terceira pergunta tenta 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 possibilitam que ferramentas, como recomendadores, os tratem de forma diferente.

Já a quarta pergunta verifica 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 - as conhecidas. 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.

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. No nosso estudo, as novidades e as conhecidas são artistas. Com os dados históricos, conseguimos identificar as novidades, as conhecidas e as preferências dos ouvintes. Com os metadados, conseguimos identificar os seus aspectos - familiaridade e popularidade.

Descobrimos que não há uma correlação geral entre a familiaridade ou popularidade e as preferências das novidades pelo ouvinte. 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 identificar 7 grupos de ouvintes baseados na em 5 características: relação entre a familiaridade e as preferências, a relação entre a popularidade 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 das novidades e conhecidas, 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 as conhecidas que para as novidades.

Com nossos resultados, indicamos a necessidade do tratamento multi-dimensional das novidades. Claramente os ouvintes possuem preferências diferentes para diferentes aspectos das novidades. Isso pode ajudar no aperfeiçoamento de sistemas de recomendação de novidades musicais. Seria necessário um recomendador personalizado, não apenas baseado no histórico de execução do ouvinte, mas também nas preferências ou pela familiaridade ou pela popularidade das novidades.

A descoberta dos grupos pode permitir que o os sistemas desenvolvam soluções diferentes para usuários diferentes. Por exemplo, construir interfaces diferentes para cada grupo, algoritmos diferentes de recomendação ou até direcionamento diferente de notícias.

Por fim, o comportamento diferenciado do ouvinte para os aspectos das novidades e para os aspectos das conhecidas corrobora que a novidade é um âmbito especial no consumo de música, precisando ser tratada de maneira específica.

1.1 Organização do documento

/* Será escrita no final (lma) */

Capítulo 2

Novidades e descobertas no âmbito musical

/* RASCUNHO (lma) */

Trabalhos sobre importância da diversidade e da novidade

Soluções para novidade > baseadas na familiaridade

Soluções para novidade > baseadas na popularidade

Contribuições

Capítulo 3

Características da novidade

/* RASCUNHO (lma) */

Definição de novidade / conhecida

Popularidade

Familiaridade

Preferência

Capítulo 4

Dados

Após a descrição das características das novidades (e conhecidas) utilizadas no nosso estudo, esta seção irá descrever 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 as novidades, as conhecidas, e as preferências dos ouvintes por ambas. Já os metadados foram utilizados para identificar os aspectos das novidades/conhecidas - a familiaridade e a popularidade. Os dados foram coletados da plataforma do LAST.FM, a partir do serviço de API fornecido pelo site. [1]

4.1 Last.FM

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

O Last.FM fornece uma API que permite o acesso a dados presentes no site. É possível coletar informações dos usuários, histórico de escuta dos usuários e informações sobre as músicas / álbuns / artistas. Para nossos experimentos, nós coletamos 2 tipos de dados: o

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

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. Já os artistas que o usuário nunca escutou anteriormente são as novidades, enquanto os que ele já escutou são as conhecidas.

4.2 Ouvinte

Com o intuito de estudar os artistas escutados pelos ouvintes, foram coletados dados acerca de um conjunto de usuários do LAST.FM. A coleta deste conjunto de usuários foi feita a partir do procedimento de *SnowBall Sampling* [1], 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. Esta coleta resultado num conjunto de 100 mil usuários.

Após a definição do conjunto de sujeitos, 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 na subseção 4.2.1.

4.2.1 Timeline

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 (HI)* do sujeito e o *período de experimento (PE)*. O HI contempla o perído desde a primeira música que o sujeito no LAST.FM até Agosto de 2012, enquanto o PE engloba o período entre Agosto de 2012 e Agosto de 2013 (um ano no total). Além dessa divisão, especificamos os seis primeiros meses do PE como *período de observação (PO)*.

Com esta divisão, foram identificadas quais as novidades escutadas pelo usuário. Os artistas escutados pelo usuário no PO que não foram escutados no HI são consideradas novidades. Não consideramos o PE todo para evitar viés no cálculo das características das novidades. Uma novidade que foi exposta no começo do PE tem maior probabilidade de ser escutada mais vezes que uma novidade que foi exposta no final do PE. Assim, identificamos como novidades os artistas escutados no PO, e levamos em conta as métricas referente a elas

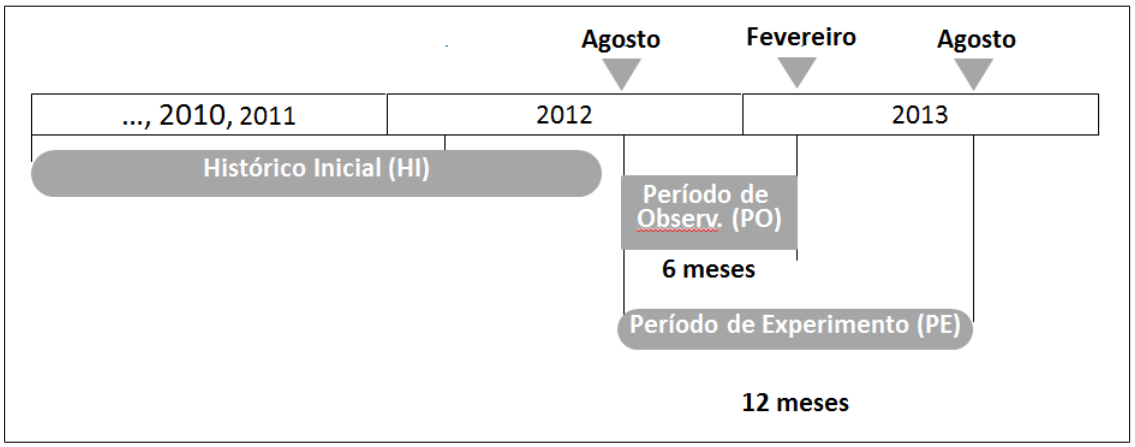

Figura 4.1: Timeline

durante todo o PE.

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*. Para o PE, 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. A segunda parte consiste nos artistas escutados em cada semana deste período, junto com o número de execuções em cada semana. O método utilizado foi o *getWeeklyArtistChart*.

Com os artistas do Histórico Total e os artistas do PE, foi encontrado os artistas do HI (complemento dos dois primeiros subconjuntos). Com os artistas do HI em mãos, finalmente foram identificadas as novidades - os artistas do PE que não estão no HI do usuário. Os dados do PE foram coletados por semana para, além de ter o número de execuções das músicas de cada artista, obter o número de semanas que o usuário escutou cada artista.

Após a coleta e definição de cada período, foi realizado uma filtragem nos dados, que será descrita no capitulo 4.2.3.

4.2.3 Filtros

Como o LAST.FM é uma rede social, formada por diferentes tipos de usuários, com diferentes hábitos musicais, foi preciso fazer 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 altos hábitos de escuta no período de HI.

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

2. Possuam altos hábitos de escuta no PE.

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

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

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

4. Possuam números irrealistas de execuções: foi detectado que alguns usuários possuiam um número muito grande de execuções musicais. Alguns, por exemplo, tiveram uma média de mais de uma música por minuto, 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 um *Scrobble*.

   Filtro: Exclusão de usuários que tiveram uma média de execuções maior que 16 horas de execuções por dia, no PE. Como as pessoas dormem em média 8 horas por dia, um ouvinte que passe o dia todo, enquanto acordado, escutando música, 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 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.

Exclusão de usuários que não escutaram nenhum artista mais de 15x na semana, em mais de 1/4 das semanas do período de observação.

O processo de filtragem resultou em um sample de 11.993 sujeitos.

4.3 Artista

Com o intuito de calcular os aspectos das novidades, onde as novidades são artistas escutados pelo usuário, foram coletados dois tipos de metadados referentes aos artistas: as *tags* que descrevem o artista e a popularidade do artista no LAST.FM. O primeiro foi utilizado para calcular a familiaridade de uma novidade para um usuário, como também para construir um modelo de perfil do usuário e calcular a ecleticidade do mesmo. O segundo foi utilizado para representar a popularidade da novidade. Todos os metadados foram coletados da API do Last.fm no dia 01 de Setembro de 2013.

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), mood (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* das músicas. Ao longo do texto, a palavra *descritor* será utilizada como um termo que descreve uma música / artista.

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. Formalmente, seja A o conjunto de artistas, e T o conjunto de tags. Seja fa : $A \times T \rightarrow R$ uma função que denote a frequência absoluta que uma tag $t \in T$ foi atribuída a um artista $a \in A$. A valor normalizado da tag t, representado pela função $f : A \times T \rightarrow R$ é representado pela equação 4.1.

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

A tabela 4.1 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*.

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

Tabela 4.1: Tags do artista Michael Jackson, junto com o valor normalizado de cada uma.

Como as tags são associadas pelos usuários do Last.FM, há problemas relacionadas a esse processo **/* REFERENCIA (lma) */**. 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 no máximo quatro tags que tenham valor mínimo de 30 (onde 100 é a tag mais atribuída àquele artista). Além disso, foram eliminadas manualmente as tags com conotação pessoal, como *seen live* (vi ao vivo), *favorite* (favorito).

Sobre a popularidade do artista, foram coletados o número de usuários do LAST.FM que escutaram cada artista. Como dito anteriormente, esses valores foram coletado no dia 01 de Setembro de 2013. Como o valor da popularidade pode atingir ordens bem diferentes, com o intuito de estudar a magnitude de cada valor, foram utilizados nos cálculos o logaritmo na base 10 do valor da popularidade. A tabela 4.2 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 4.2: Número de ouvintes (popularidade) de alguns artistas no LAST.FM

Capítulo 5

Modelos

Após a coleta e filtragem dos dados, utilizamos 3 conceitos para representar os artistas escutados pelos sujeitos:

1. Modelo de perfil do ouvinte: foi construindo o modelo do perfil do ouvinte, com o intuito de: gerar uma representação visual do que foi escutado pelo ouvinte; participar no cálculo da familiaridade de um artista para um ouvinte; e gerar a métrica de ecleticidade, que foi utilizada na criação dos grupos de ouvintes (capítulo 7).

2. Características das novidades: foram modeladas as características da novidade a serem utilizadas nos experimentos - familiaridade e popularidade.

3. Métricas de relevância: por fim, foram modeladas duas métricas que refletem a preferência do ouvinte para um artista, ou a relevância deste artista para o ouvinte, durante um período de tempo - a atenção total e o período de atenção.

5.1 Perfil musical do ouvinte

Um perfil de um usuário é um modelo, geralmente apresentado em forma de grafo ou tabela, 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 escuta.

A necessidade de uma representação do perfil musical do ouvinte surgiu primeiramente para calcular a familiaridade de um artista para o ouvinte. Como definida no capítulo 3, a

familiaridade de um artista para um ouvinte está relacionada com a semelhança dos descritores do artista com os descritores dos artistas escutados pelo ouvinte, junto com a influência destes descritores no histórico musical do ouvinte. Para realizar o cálculo, é necessário um modelo do perfil do ouvinte que sumarize os descritores das músicas por ele escutadas no decorrer do tempo, para comparar com os descritores da novidade escutada.

Uma forma de representação informal do perfil de uma pessoa seria o conjunto de gêneros de artistas que essa pessoa escuta/escutou. Tipicamente, ao perguntar a ouvintes qual seu perfil musical, respostas como estas surgem: "Meu perfil é formado por artistas de Rock", "Meu perfil é formado por artista de Forró e artistas de Pagode". Uma representação formal desse conceito se adequaria à nossa necessidade, pois gêneros musicais também são considerados como descritores.

Assim, o nosso modelo de perfil de ouvinte é formado pelo conjunto de clusters de artistas mais representativos do histórico do ouvinte, onde cada cluster é formado por artistas semelhantes de acordo com os seus descritores. Com isso, um sujeito pode ter um perfil formado por um cluster de artistas de Rock, um cluster de artistas de Samba, um cluster de artistas de Forró, e assim por diante. Os artistas do histórico musical utilizados para a construção do perfil foram os que não são novidades e que possuíssem número de execução no histórico do usuário maior que a média do número de execuções total dos artistas do histórico do usuário.

Para a construção deste conjunto de clusters, foi utilizado um algoritmo de clusterização hierárquico aglomerativo **/* REFERENCIA (lma) */**. Este tipo de algoritmo inicializa cada elemento (em nosso caso cada artista) em um cluster, e a cada passo, ele une os dois clusters mais próximos (similares). Desta maneira, é necessário definir uma medida de distância, ou dissimilaridade, entre os clusters.

Na maior parte dos métodos utilizados na clusterização hierárquica aglomerativa, a medida de distância entre clusters pode ser gerada a partir de uma métrica de distância entre os pares de elementos e um critério de união que especifica quais clusters unir em cada passo, em função desta distância. Então, foi definido como métrica de distância entre pares de artistas o complemento da similaridade do cosseno entre os vetores de tags dos artistas e como critério de união (*linkage criterion*) a clusterização de união pela média (o *average linkage clustering*).

Para calcular a similaridade do cosseno**/* REFERENCIA (lma) */**, os artistas foram representados por vetores, onde o vetor é formado pelas tags atribuídas a cada artista (e que passaram pelo processo de filtragem). Já o valor de cada tag é o valor normalizado do número de vezes que a tag foi atribuída ao artista. Tanto a filtragem quanto valor de cada tag foram descritos na seção 4.3. Assim, seja A o conjunto de artistas, e T o conjunto de tags. Seja $f : A \times T \rightarrow R$ a função que denote a frequência que uma tag $t \in T$ foi atribuída a um artista $a \in A$. O vetor que representa um artista $a \in A$ está definida na equação 5.1.

$$\vec{a} := (f(a, t_1), f(a, t_2), \ldots, f(a, t_{|T|})) \tag{5.1}$$

A similaridade do cosseno entre dois artistas é definida pela equação 5.2. Como o algoritmo aglomerativo hierárquico requer uma medida de distância, e não de similaridade, foi calculado o complemento da similaridade do cosseno (5.3).

$$cos(\vec{a}, \vec{a}') := \frac{\langle \vec{a}, \vec{a}' \rangle}{\|\vec{a}\| \|\vec{a}'\|} \tag{5.2}$$

$$dis(\vec{a}, \vec{a}') := 1 - cos(\vec{a}, \vec{a}') \tag{5.3}$$

O *average linkage clustering* **/* REFERENCIA (lma) */** é um método de união de clusters baseado na média das distâncias entre cada par de elementos de cada cluster. A distância entre dois clusters é definida pela equação 5.4. Seja X e Y clusters, $x \in X$ um artista do cluster X e $y \in Y$ um artista do cluster Y. A distância $d(X, Y)$ entre os clusters X e Y é definida pela média das distâncias de todos os pares $x \in X$ e $y \in Y$.

$$d(X, Y) := \frac{1}{|X|\,|Y|} \sum_{x \in X} \sum_{y \in Y} dis(x, y) \tag{5.4}$$

Após a definição da distância entre clusters, o algoritmo de clusterização foi aplicado. Como é um método aglomerativo hierárquico, o algoritmo inicia cada artista dentro de um cluster separado. Em cada etapa os clusters mais próximos vão sendo aglutinados, até chegar em 1 cluster com todos os artistas. Para selecionar o número de clusters de um ouvinte, o algoritmo foi interrompido no momento em que a distância mínima entre 2 clusters fosse igual a 0.30.

Os perfis obtidos tiveram média de 33,28 clusters (onde cada cluster possui pelo menos 2 artistas) e desvio padrão de 20,3. A figura **/* FIGURA (lma) */** representa o histograma do número de clusters. Um exemplo de perfil está representado na figura **/* FIGURA (lma) */**. Como comparação, a figura **/* FIGURA (lma) */** possui o grupo de artistas utilizados na clusterização.

A criação do perfil, além de auxiliar na visualização do gosto musical do usuário, evidenciado nas figuras **/* FIGURAS (lma) */**, e do cálculo da familiaridade (seção 5.2.1), faz parte do cálculo da ecleticidade.

5.1.1 Ecleticidade

A ecleticidade representa o quão eclético musicalmente um ouvinte é - o quão diferente são os 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 7.

Inicialmente foi pensado em utilizar o número de clusters do perfil do ouvinte como critério de ecleticidade. Quanto mais clusters o ouvinte possuísse no perfil, mais eclético ele seria. Porém, dois ouvintes podem possuir o mesmo número de cluster mas um ouvinte pode possuir no perfil clusters mais similares (como um cluster de indie rock e outro de british rock) e outro possuir menos similares (como um cluster de forró e outro de indie rock).

Uma alternativa a essa abordagem seria contabilizar o quanto de diferença ou diversidade cada cluster adiciona ao perfil. Quanto mais diversidade cada cluster proporcionar, mais eclético o ouvinte é. Para isso aplicamos um algoritmo de clusterização hierárquico dos clusters do perfil, e a cada passo calculamos a distância entre os dois clusters que foram unidos. Assim, a ecleticidade representa a soma de todas estas distâncias. Formalmente, seja $P := \{C_1, ..., C_n\}$ o perfil do ouvinte, formado pelos clusters do perfil C_i. Seja $X^{(p)} := \{X_1^{(p)}, ..., X_n^{(p)}\}$ o conjunto de clusters no passo p do algoritmo hierárquico, onde $X^{(1)} = P$ e $X^{(j)}$, onde $j > 1$, clusters do algoritmo hierárquico criados a partir do conjunto inicial $X^{(1)}$. Além disso, seja $d(X_k^{(p)}, X_l^{(p)})$ a distância entre os clusters $X_k^{(p)} \in X^{(p)}$ e $X_l^{(p)} \in X^{(p)}$. Então, a ecleticidade do ouvinte com perfil P está representada pela equação 5.5.

$$e(P) = \sum_{i=1}^{|x^{(p)}|} min(d(X_k^{(p)}, X_l^{(p)})) \tag{5.5}$$

A distância $d(X_k^{(p)}, X_l^{(p)})$ entre os clusters do algoritmo hierárquico foi calculada utilizando o *average linkage method* (equação 5.4). Como o *average linkage method* depende da distância entre cada par de elemento (onde cada elemento é um cluster $C_i \in P$), foi definido como distância entre dois clusters $C_i \in P$ e $C_j \in P$ o completo da similaridade do cosseno (equação 5.3) entre os vetores c_i e c_j que representam o centróide de C_i e C_j.

5.2 Características das novidades

Para caracterizar as novidades multidimensionalmente, utilizamos dois aspectos: a familiaridade e a popularidade. Esta seção descreve o cálculo de cada aspecto.

5.2.1 Familiaridade

Como definido no capítulo 3, a familiaridade de um artista a para um ouvinte o reflete o quanto este ouvinte conhece os descritores do artista a. A familiaridade está relacionada com a similaridade entre os descritores do artista e os descritores dos artistas do perfil do ouvinte, junto com a influência desses artistas.

Formalmente, seja $P := \{C_1, ..., C_n\}$ o perfil do ouvinte, formado pelos clusters de artistas C_i. Seja $\vec{c_i}$ o centróide do cluster C_i, e p_i a influência do cluster C_i no perfil do ouvinte, representado pela proporção de todas as execuções dos artistas de C_i pelo ouvinte. Assim, a familiaridade entre um artista a e o perfil do ouvinte P é o valor máximo da similaridade entre o artista a e algum cluster C_i multiplicado pela influência p_i de C_i:

$$fam(a, P) = max(cos(\vec{a}, \vec{c_i}) \times p_i) \tag{5.6}$$

~~A influência p_i de C_i é representado pela média do total de execuções das músicas dos artistas presentes no cluster. Quanto mais vezes os artistas do cluster i foram escutados, mais influentes os descritores deste cluster são para o ouvinte.~~

5.2.2 Popularidade

O segundo aspecto da novidade estudado foi a popularidade. Para calcular a popularidade, utilizamos o logaritmo na base 10 da número de ouvintes do artista no LAST.FM. O logaritmo foi utilizado pois a distribuição da popularidade dos artistas é enviesada (5.1).

Figura 5.1: Curva de popularidade dos artistas do LAST.FM. A curva é enviesada.

5.3 Preferências

Para mensurar o quanto o ouvinte preferiu a novidade, foram utilizados duas métricas: a atenção total e o período de atenção. Por causa que novidades podem ser descobertas em todo o PE, alguns destes artistas possuem uma janela de tempo no experimento menor (artistas escutadas no final do PE). Para contornar esse problema, utilizamos duas soluções. Primeiro, utilizamos como denominador no cálculo das métricas o número de semanas período JT, que representa a janela de tempo de exposição à novidade, que vai da primeira semana que foi escutada a novidade até o fim do PE. Segundo, como mencionado na seção 4.2.1, apenas as novidades descobertas no PO foram consideradas na análise, mas todo o PE 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 PE. A atenção total do ouvinte para o artista é representada pelo total de número de execuções de músicas do artista que ele escutou no tempo JT, dividido pelo número de semanas de sua JT.

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 JT.

Capítulo 6

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.

6.1 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 relação geral entre algum aspecto da novidade e as preferências dos ouvintes? Para responder essa pergunta, calculamos a correlação entre algum aspecto da novidade - familiaridade e popularidade - e alguma métrica de preferência - atenção total e período de atenção, de todas as novidades de todos os ouvintes juntas. Caso algum par de correlação seja alta, podemos afirmar que, no geral, os ouvintes preferem novidades deste aspecto.

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/descreve. 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	PdA	F	P
Atenção Total (AT)	0.71	0.06	0.06
Período de atenção (PdA)	-	0.08	0.07
Familiarity (F)	-	-	0.08
Popularidade (P)	-	-	-

Tabela 6.1: Correlação (Coeficiente de Spearman) entre aspectos da novidade e preferências, analisando todas as novidades juntas

A tabela 6.1 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, os ouvintes em geral não preferem novidades familiares, ou no geral não preferem novidades não-familiares. Nós levantamos duas hipóteses para explicar esse resultado:

1. Diferentes ouvintes possuem diferentes preferências para os aspectos das novidades. Nesta hipótese, alegamos que diferentes ouvintes possuem diferentes preferências musicais. Assim, existem ouvintes que preferem novidades populares, outros preferem novidades não populares, etc. Colocando todos estes ouvintes juntos, a correlação geral vai ser próxima a zero.

2. Individualmente, os ouvintes não preferem um aspecto a outro das novidades. Cada ouvinte pode preferir, por exemplo, tanto novidades familiares quanto não familiares, fazendo com que a correlação entre a preferência e o aspecto das novidades seja próxima a zero.

6.2 Preferências individuais

O resultado da primeira pergunta de pesquisa e estas hipóteses 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 **/* FIGURA (lma) */** 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 um comportamento comum frente as 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. Isto indica 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 capítulo próximo capítulo.

Capítulo 7

Grupos de usuários para diferentes aspectos de novidade

De acordo com os resultados do capítulo 6, 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 realizada uma clusterização nos dados que caracterizam os sujeitos. Esta análise será descrita neste capítulo.

7.1 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 6). 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, enquanto os outros resultados estão no apêndice **/* X (lma) */**.

7.2 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. 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 familiaridade e a atenção total*
    - *Correlação entre popularidade e a atenção total*

    **Como há uma correlação forte entre a atenção total e o período de atenção, relatado na tabela 6.1, decidimos utilizar na clusterização apenas as correlações que envolvem a atenção total*

2. **Relacionadas com os hábitos musicais**: Métricas que caracterizam os hábitos musicais do sujeito e que estão relacionadas com 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 5.1.1. 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 está relacionada com a popularidade das novidades escutadas pelo ouvinte.
    - *Proporção de novidades escutadas pelo ouvinte no PO* Com esta métrica pode-se identificar se o ouvinte possui o hábito de escutar muitas ou poucas novidades.

7.3 Algoritmo de agrupamento

Para fazer o agrupamento dos sujeitos, foi utilizado o algoritmo de agrupamento aglomerativo hierárquico. Como descrito na seção 5.1, 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 7.2. 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, seja 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 $disSuj(x, y)$ dos vetores que os representam é dado pela equação 7.1. Como critério de união, foi utilizado o método Ward **/* REFERENCIA (lma) */**.

$$disSuj(x, y) := \sqrt{\sum_{i=1}^{5}(x_i - y_i)^2} \tag{7.1}$$

7.4 Escolha do número de clusters

O método de clusterização hierárquica não expõe explicitamente o número de clusters resultantes. A cada passo, o algoritmo une dois clusters, até que todos os sujeitos estejam em um só cluster. Uma abordagem para definir a melhor configuração de número de clusters é o método do joelho **/* REF (lma) */**, ao plotar um gráfico onde o eixo X é o número de clusters e o Y um critério de avaliação. A figura 7.1 mostra a distância média dentro dos clusters para cada configuração de número de clusters. O método de joelho determina escolher uma configuração de clusters que não adicionem muita heterogeneidade, evidenciado a partir da curvatura máxima do gráfico (joelho). Pela figura, a distância média dentro dos clusters começa a aumentar vertiginosamente nas configurações com menos de 7 grupos. Assim, analisamos as configuração com 6, 7 e 8 clusters.

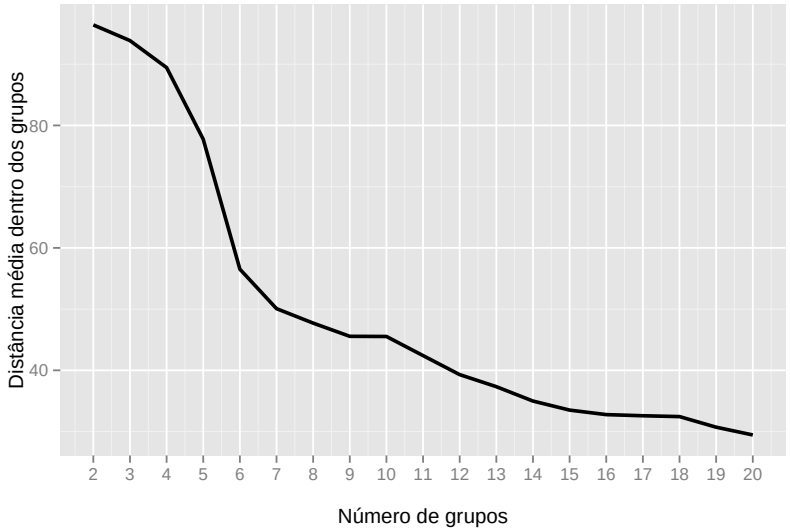

Figura 7.1: Número de cluster X Distância média dentro dos clusters. O joelho do gráfico está em torno da configuração com 7 clusters.

Comparando os centróides dos clusters para as configurações 6, 7 e 8, os centróides dos clusters 7 e 8 são semelhantes. **/* figura Z (lma) */** Desta maneira, escolhemos a solução com 7 grupos a mais representativa para nossos estudos.

7.5 Clusters

Após a escolha de 7 grupos de ouvintes, utilizamos o centróide de cada grupo para analisar as suas principais características. A figura 7.2 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

   (a) Averso a coisas 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.

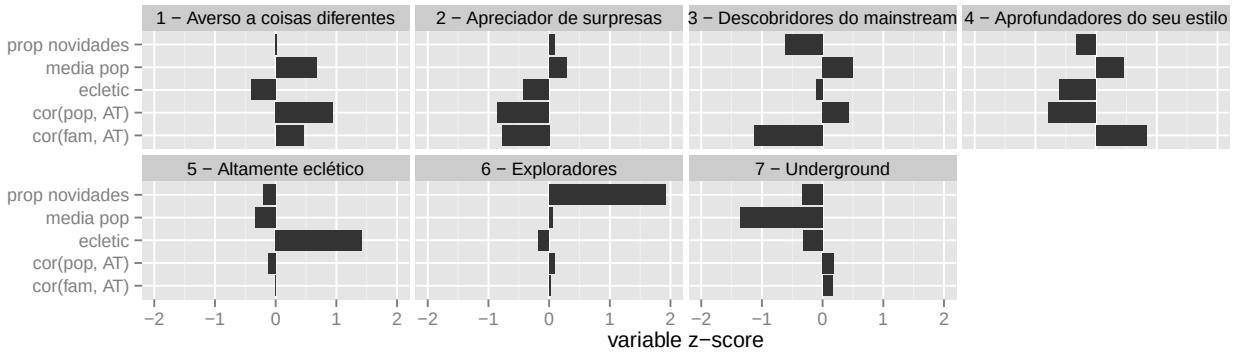

Figura 7.2: Centróides dos 7 grupos encontrados na análise. As métricas estão normalizadas pelo z-score, onde zero representa a média de todos os ouvintes, e a unidade de variação é um desvio padrão, para cada métrica. No eixo vertical, *fam* significa a familiaridade, *pop* significa popularidade, *AT* significa atenção total.

(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) Aprofundadores do seu estilo [total de ouvintes: 1467 (14%)]: Ouvintes que preferem novidades pouco populares e familiares, além de possuir pouca ecleticidade.

2. Grupos marcados pelas características dos hábitos musicais

(a) Altamente eclético (2) [total de ouvintes: 2456 (21%)]: Maior grupo de todos, formado por ouvintes que possuem alta ecleticidade

(b) Exploradores (3) [total de ouvintes: 1047 (9%)]: Ouvintes que possuem alta proporção de novidades escutadas durante o PO

(c) Underground (4) [total de ouvintes: 1281 (11%)]: Ouvintes com hábito musical marcado por artistas pouco populares.

7.6 Explicação dos clusters

Encontramos 4 grupos que possuem características principais pelas 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 coisas 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.

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 coisas 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 preferindo novidades não familiares e não populares. Eles preferem surpresas, artistas diferentes do que já escutaram.

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 *Aprofundadores do 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.

7.7 Discussão dos resultados

Capítulo 8

Comparação das novidades com as conhecidas

Capítulo 9

Conclusão

Capítulo 10

Trabalhos Futuros

Bibliografia

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

[2] S. R. Maddi. *Theories of cognitive consistency*, chapter The pursuit of consistency and variety. Rand McNally, Chicago, 1968.

Apêndice A

Apêndices