Análise de Coocorrência de Itens Executados em Programações de Rádios

Alternative Title: Co-occurrence Analysis for Items Played in Radio Stations Programs

Alexandre P. Norberto
alexandre.pnorb@gmail.com

Luiz H. C. Merschmann
luizhenrique@iceb.ufop.br

Amanda S. Nascimento
amanda.nascimento@gmail.com

Álvaro R. Pereira Jr.
alvaro@iceb.ufop.br

Departamento de Computação (DECOM)
Universidade Federal de Ouro Preto (UFOP)
Ouro Preto - MG - Brasil

RESUMO

Sistemas de recomendação de músicas, como a Last.fm, utilizam bases de conhecimento coletivo, contendo classificações e históricos de uso, para recomendar itens (como artistas e músicas) considerados similares entre si. Tendo como referência esse conhecimento gerado pelos usuários da Last.fm, este trabalho investiga a coocorrência de itens nas programações de estações de rádio, isto é, se as listas de reprodução das rádios representam conjuntos coesos de faixas relacionadas. Nesse sentido, foi elaborada uma metodologia de análise, na qual são extraídos conjuntos de itens frequentes contendo artistas, músicas e gêneros musicais que coocorrem nas programações de rádios, analisando, em seguida, a similaridade entre os itens dos conjuntos gerados, utilizando como referência as listas de itens similares extraídas da Last.fm. Os resultados experimentais deste trabalho revelam que quanto mais rigorosos são os filtros aplicados para se definir os conjuntos de itens frequentes, mais itens presentes nas listas de similaridade da Last.fm são encontrados, através da correlação descoberta pela programação das rádios. No entanto, o estudo também revela que novas correlações entre itens, ainda não classificadas na Last.fm, podem ser descobertas através das programações das rádios, o que evidencia que até mesmo sistemas utilizados em larga escala e de forma colaborativa, como Last.fm, não são completos no que diz respeito à caracterização dos dados para abstração do conceito de similaridade.

Palavras-Chave

Músicas, Mineração de Dados, Análise de Similaridade

SBSI 2016, May 17th-20th, 2016, Florianópolis, Santa Catarina, Brazil

ABSTRACT

Music recommendation systems such as Last.fm use collective knowledge bases containing ratings and historical use, in order to recommend items (such as artists and songs) considered similar to each other. With reference to the knowledge generated by Last.fm users, this work investigates the co-occurrence of items in radio station programs, that is, whether the radio playlists represent cohesive sets of related items. In this sense, we developed an analysis methodology in which frequent item sets containing artists, songs, and musical genres that co-occur within radio programs are extracted, in order to analyze the similarity among discovered item sets, using as reference the similarity lists from Last.fm. Experimental results show that the more strict the filters applied to define the set of frequent items are, the more items present in the Last.fm similarity lists are found. On the other hand, the study also reveals that new correlations between items, even though not classified by Last.fm, can be discovered, showing that even large-scale collaborative systems as Last.fm are not complete regarding data characterization for extracting abstract similarity concepts.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms

Experimentation

Keywords

Music, Data Mining, Similarity Analysis

1. INTRODUÇÃO

Em nosso dia a dia frequentemente vivenciamos situações em que é importante conhecer a similaridade entre os elementos presentes a nossa volta. Quando gostamos de algo, naturalmente escolhemos por outros itens que são relacionados ao que gostamos. Por exemplo, uma pessoa que gosta de dançar, frequentemente escolhe ir a locais com música

e espaço para dança. Já uma pessoa que é sensível à música alta, deve preferir frequentar locais mais silenciosos. De forma intuitiva, natural e muitas vezes inconsciente, fazemos associações de similaridade ou dissimilaridade entre os elementos do mundo que estão a nossa volta.

Ao representar um dado domínio na construção de sistemas de informação, projetistas muitas vezes precisam, de forma explícita e lógica, representar relações de similaridade entre itens do domínio que estão presentes no sistema, muitas vezes tentando imitar o mundo real. Por exemplo, um sistema efetivo de recomendação de bares e restaurantes precisa ser capaz de capturar o conceito abstrato de que o usuário em questão, digamos, não gosta de lugares com música alta e que, por outro lado, um dado restaurante tem um ambiente tranquilo e aconchegante. A partir da caracterização tanto do usuário quanto do restaurante em questão, tal sistema deve finalmente ser capaz de relacionar estes dois elementos, de forma a recomendar para o usuário o restaurante com ambiente tranquilo. Este relacionamento só pode ser feito de fato se o sistema possuir meios de estabelecer o conceito de similaridade entre os itens do domínio, o que muitas vezes é muito difícil, ou até mesmo impossível de ser feito com uma acurácia aceitável para a aplicacão em foco.

Um domínio cujo conceito de similaridade entre itens é ponto chave para a qualidade da aplicação é representado pelos sistemas de recomendação de música como Spotify[1], Deezer[2], Last.fm[3], dentre muitos outros disponíveis atualmente. Nesses sistemas, se o usuário realiza uma busca por um dado nome de artista ou gênero musical, o sistema deve apontar itens similares para compor uma *playlist* com artistas variados, porém relacionados de alguma forma com o item buscado. No desenvolvimento de sistemas de recomendação, ser capaz de capturar o conceito de similaridade pode representar um grande desafio, seja usando de dados do próprio sistema, ou mesmo de dados disponíveis de alguma forma na *web*.

Neste trabalho, investiga-se a qualidade com que se é possível extrair relações de similaridade (isto é, correlações), entre itens que são executados frequentemente em uma mesma estação de rádio. A hipótese central é que, quanto mais dois itens i e j se repetem na programação de uma dada rádio, e quanto mais rádios apresentam em sua programação os itens i e j, mais similares são os itens i e j. Apesar da dificuldade em se avaliar a qualidade da correlação entre itens, o presente trabalho apresenta uma proposta de métrica de avaliação que parece ser bem efetiva. A métrica proposta faz uso das relações de similaridade disponibilizadas pela Last.fm.

Em suma, no presente trabalho é investigado se a coocorrência de itens como artistas, músicas e gêneros musicais, executados em programações de estações de rádio, indica a existência de correlação entre os itens. A motivação de tal investigação parte do pressuposto de que existe uma lógica nas programações das rádios e, normalmente, as músicas não são escolhidas de forma aleatória, pois os locutores ou DJs, em geral, possuem um grande conhecimento sobre os conteúdos que eles executam, escolhendo cuidadosamente o sequenciamento dos itens a partir do seu próprio conceito intuitivo e abstrato da similaridade entre os itens, a fim de evitar a quebra de identidade do programa [12].

Para analisar a correlação entre os itens musicais das rádios, utiliza-se como referência a base de conhecimento coletivo gerada pela Last.fm. Em geral, sistemas de recomendação de músicas utilizam abordagens como Filtragem Colaborativa (*Collaborative Filtering*) e técnicas baseadas em conteúdo (*Content-Based Techniques*) [16], explorando o grande volume de informações referentes aos hábitos e classificações dos usuários no sistema para fazer o processo de recomendação de músicas e artistas. Esse é o caso do serviço da Last.fm, que além de ser um sistema de recomendação de músicas é também uma rede social [2]. Atualmente, a Last.fm continua sendo uma das principais plataformas sociais de músicas, com mais de 40 milhões de usuários ao redor do mundo.

Os usuários da Last.fm trocam, de forma colaborativa, informações sobre músicas e artistas, que são classificados de acordo com seus gêneros musicais e agrupados conforme similaridades encontradas pelo próprio sistema. Além disso, as informações contidas no perfil de cada usuário da Last.fm sobre as músicas que foram ouvidas por eles são mantidas no sistema. Dessa forma, esse conhecimento coletivo gerado pelos próprios usuários permite que a Last.fm consiga inferir similaridade entre itens (como artistas, músicas e gêneros musicais). Essas listas de itens similares produzidas pela Last.fm são importantes fontes de informação, sendo utilizadas também em outros trabalhos que envolvem o estudo de similaridade entre itens musicais, como os trabalhos [4, 3].

Para a realização do estudo, numa primeira etapa, técnicas de mineração de regras de associação [1] foram utilizadas para extração de conjuntos de itens frequentes a partir de bases de dados que foram construídas utilizando informações da programação de estações de rádio. Tais conjuntos contêm itens que coocorrem nas programações das rádios com uma determinada frequencia. Em seguida, foi verificado o grau de similaridade entre os itens contidos nos conjuntos frequentes minerados e aqueles contidos nas listas de itens similares extraídas da Last.fm. Com isso, foi possível verificar se as listas de reprodução das rádios representam conjuntos coesos de faixas relacionadas.

O restante deste trabalho está organizado da seguinte maneira. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta a metodologia proposta neste trabalho. Em seguida, a Seção 4 apresenta e discute os resultados experimentais. Por fim, na Seção 5 são apresentadas as conclusões e sugestões de trabalhos futuros.

2. TRABALHOS RELACIONADOS

Existem diversos trabalhos na literatura que envolvem a aplicação de técnicas de mineração de dados para a realização de análises no contexto musical.

O trabalho [11] demonstra como a mineração de regras de associação pode ser aplicada para se descobrir relações entre duas ontologias (gênero e região) da música *folk*. Dadas as associações entre gêneros e regiões, o método identifica combinações entre conteúdo-região, conteúdo-gênero e região-gênero. Esse trabalho representa uma contribuição para a área de Recuperação de Informação Musical em termos de tarefa de recuperação (descoberta de associações entre gêneros da música *folk* e sua distribuição geográfica), além de permitir uma combinação sistemática de modelos de regras e métricas de avaliação para distinguir categorias

[1]https://www.spotify.com/
[2]https://www.deezer.com/
[3]http://www.last.fm/

de associações.

O trabalho [8] apresenta uma abordagem não supervisionada para agrupamento de músicas a partir de características extraídas de seus sinais de áudio. Assim, graças à forte correlação entre as músicas contidas em um mesmo grupo, a informação contida em cada grupo formado pode servir para auxiliar na identificação de características importantes nos sinais de áudio das músicas de um dado grupo. O resultado de uma avaliação empírica sobre um conjunto de dados de 91 músicas *pop* apresentou uma pureza média do cluster de 57,3%.

Os autores do trabalho [9] utilizam os algoritmos Apriori e *Sequential Pattern Mining* (SPAM) para buscar padrões sequenciais frequentes em uma base de dados (Yahoo! *Music* [6]) contendo bilhões de classificações de músicas definidas por usuários. Quando analisados corretamente, esses dados podem revelar informações de como a popularidade de músicas, álbuns e artistas varia ao decorrer do tempo. Também a partir dos padrões descobertos, é possível saber quais sequências de músicas são frequentemente ouvidas. Os resultados experimentais revelam que o algoritmo SPAM possui uma boa performance para grandes bases de dados como a Yahoo! *Music*.

Por fim, o trabalho [12] apresenta uma abordagem para se extrair automaticamente similaridades entre músicas ou artistas, baseando-se na coocorrência desses itens em diferentes fontes, como a *playlist* de uma estação de rádio francesa (Fip - Radio France) e em álbuns de coletâneas musicais (utilizando a CDDB, que é uma base de dados de álbuns musicais disponível na *web*). Para isso, a técnica utilizada consiste em construir uma matriz de distâncias entre os itens, baseando-se na coocorrência dos mesmos. Através dessa matriz, os itens foram agrupados, tendo como referência a distância entre eles. Os experimentos foram conduzidos sobre três tipos de bases de dados definidas: uma bem pequena contendo 12 títulos de músicas, outra contendo 80 títulos de músicas tocadas frequentemente na programação da rádio selecionada (Fip), e uma base construída com 100 artistas, coletados manualmente. Para checar a consistência das similaridades geradas pelos grupos de itens, foram feitos julgamentos por cinco pessoas especialistas em música. Os resultados preliminares mostram que a técnica consegue extrair similaridades entre itens, em pequenas bases de dados.

Nota-se que a proposta do trabalho [12] é semelhante ao que é apresentado neste. No entanto, ele utiliza poucas fontes de dados para as análises, como a coleta da programação musical de apenas uma rádio. O tamanho das amostras utilizadas nos experimentos também é outro ponto a se destacar, pois amostras tão pequenas poderiam interferir negativamente nos resultados e conclusões do trabalho. Além disso, a base de dados de artistas selecionados manualmente também pode gerar resultados tendenciosos. Os resultados obtidos precisaram ser julgados manualmente, por especialistas da área musical, para avaliar as similaridades extraídas, o que parece ser outra limitação do trabalho.

3. METODOLOGIA

Nesta seção apresenta-se a metodologia elaborada para o estudo de correlação de itens, que abrange o processo de construção das bases transacionais utilizadas neste trabalho (Seção 3.1), a extração dos conjuntos de itens frequentes a partir das bases transacionais (Seção 3.2) e a comparação entre os conjuntos gerados e as listas de itens similares fornecidas pela Last.fm.

A Figura 1 ilustra o fluxo das etapas da metodologia desenvolvida. As etapas são divididas em duas fases. Na primeira fase (Preparação das bases transacionais), foram utilizadas técnicas de mineração de regras de associação para extração de conjuntos de itens frequentes (músicas, artistas e gêneros musicais) a partir de bases de transações que foram construídas utilizando-se dados da programação de estações de rádio. Na segunda fase (Comparação das listas de coocorrências), foi verificado o grau de similaridade entre os itens contidos nos conjuntos frequentes minerados e aqueles contidos nas listas de itens similares extraídas da Last.fm. Com isso, foi possível verificar se a coocorrência de itens (músicas, artistas e gêneros musicais) em programações de rádios indica a existência de correlação (similaridade) entre eles, tendo como referência a Last.fm. Mais detalhes sobre as etapas da metodologia, ilustradas na Figura 1, são descritos nas seções subsequentes.

3.1 Preparação das bases transacionais

Como mostra a Figura 1, a primeira fase da metodologia começa com a coleta de metadados de programações de estações de rádio. Este trabalho utilizou os metadados coletados a partir dos serviços do projeto Radialize[4]. É importante ressaltar que esses metadados são públicos e podem ser coletados através da extração de dados nas páginas de transmissão das rádios pela *web*, pois geralmente são informados metadados, como os nomes do artista e da música no momento em que ela está sendo executada na rádio.

Para o presente estudo, a tarefa de coleta não foi realizada de fato, uma vez que o projeto Radialize forneceu a sua base de dados que havia sido previamente coletada. Os dados fornecidos correspondem a registros de execuções musicais relativos ao período de janeiro de 2011 a abril de 2014, correspondendo a 58.247 artistas, 335.676 músicas e 2413 rádios, de um total de 18.624.318 de registros. Cada registro contém os seguintes campos: código identificador da rádio, código identificador do artista, código identificador da música, nome do artista, nome da música, *timestamp* do início da execução da música e *timestamp* do fim da execução da música.

Após a carga dos dados na base de programação das rádios, o processo se inicia com a etapa 1.1 (Geração de bases transacionais), na qual essa base é transformada em bases transacionais, de maneira que cada transação representa a programação de uma rádio, sendo que seus itens podem ser artistas, músicas ou gêneros musicais contidos na programação daquela rádio. Filtros de número de ocorrências mínimas de itens nas programações das rádios foram utilizados na geração das diferentes bases transacionais. Desse modo, a partir de um número mínimo de ocorrências definido no filtro, um item (por exemplo, artista) é incluído em uma transação apenas se ocorrer na programação da rádio esse número mínimo de vezes estabelecido no filtro. Ao longo deste trabalho serão apresentados resultados de experimentos realizados a partir de bases geradas utilizando-se diferentes valores de frequência para os filtros.

Como a base de metadados de programações de rádios não contém informações explícitas sobre os gêneros musicais presentes nas programações das rádios, essas informações foram coletadas a partir da Last.fm para se gerar as bases transacionais de gêneros musicais, como ilustra a etapa

[4]https://www.radialize.com.br

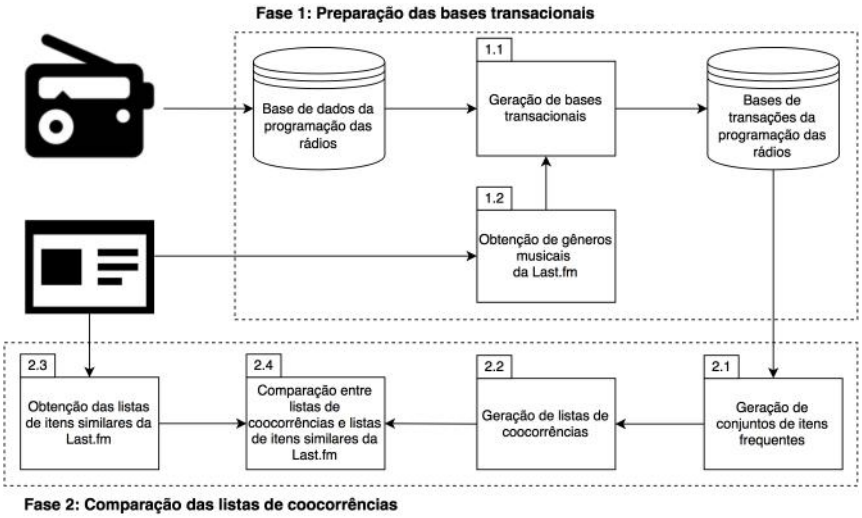

Figura 1: Etapas da metodologia elaborada

1.2 (Obtenção de gêneros musicais da Last.fm) na Figura 1. As bases transacionais de gêneros musicais foram geradas a partir das bases transacionais de artistas, sendo que cada artista presente em uma transação foi substituído pela primeira *tag* da lista de *tags* relacionadas àquele artista na Last.fm. Da lista de *tags* de cada artista, obtida da Last.fm por meio de sua API (*Application Programming Interface*), foi utilizada apenas a primeira *tag*, pois essas *tags* são definidas por usuários e aquelas mais bem classificadas [7], ou seja, que estão no topo da lista, geralmente, representam os principais gêneros musicais daquele artista.

A Tabela 1 apresenta as características das bases transacionais geradas para artistas, músicas e gêneros musicais. Essa tabela apresenta os filtros que dizem respeito ao número mínimo de ocorrências de um item nas programações das rádios (1ª coluna). A tabela apresenta também, para essas bases transacionais, o número de transações (2ª coluna), o número médio de itens das transações (3ª coluna) e o desvio padrão do número de itens das transações (4ª coluna).

Como pode ser visto, os filtros do número mínimo de ocorrências de itens foram variados (sem filtro, 3, 5 e 10). Nesse caso, quanto maior é o valor do filtro, menor é o número de rádios cuja programação atende ao critério estabelecido e, consequentemente, menor é o número total de transações da base gerada (lembrando que cada transação representa o conteúdo executado na programação de uma rádio).

Tabela 1: Bases transacionais

Artistas			
Filtro	**Nº Trans.**	**Tam. Méd.**	**Desv. Pad.**
Sem filtro	2402	351,98	259,55
3	2031	277,08	187,91
5	1775	246,11	165,59
10	1514	214,72	139,86
Músicas			
Filtro	**Nº Trans.**	**Tam. Méd.**	**Desv. Pad.**
Sem filtro	2413	895,80	822,52
3	1891	585,54	421,11
5	1632	494,98	339,46
10	1325	399,22	265,69
Gêneros Musicais			
Filtro	**Nº Trans.**	**Tam. Méd.**	**Desv. Pad.**
Sem filtro	2374	143,86	94,69
3	2001	177,24	72,24
5	1747	105,63	65,19
10	1484	92,81	56,76

3.2 Comparação das listas de coocorrências

A partir das bases transacionais de artistas, músicas e gêneros musicais, foram extraídos os conjuntos de itens frequentes, sendo a etapa 2.1 (Geração de conjuntos de itens frequentes) da metodologia, utilizando-se para isso a implementação do algoritmo Apriori [15], disponível no *software* R [13]. Embora existam soluções mais otimizadas para extração de conjuntos frequentes, o Apriori atendeu às necessidades deste estudo, pois foram extraídos apenas conjuntos frequentes de tamanho 2, que posteriormente foram processados para gerar as listas de itens (músicas, artistas e gêneros musicais) que coocorrem nas programações das rádios, tratando-se da etapa 2.2 (Geração de listas de coocorrências). Por exemplo, suponha os conjuntos frequentes {A, B}, {A, C} e {A, D}. Isso indica que, na programação das rádios, o item A coocorre com os itens B, C e D.

Foram testados os valores de suporte mínimo de 5%, 10%, 15% e 20%, ou seja, variou-se os valores de suporte começando pelo mínimo aceitável (5%) para as análises, até o máximo possível de se obter conjuntos de itens frequentes (10%, 15% e 20% para músicas, artistas e gêneros musicais, respectivamente).

Após a geração das listas de coocorrências, realiza-se a comparação das mesmas com as listas de itens similares obtidas da Last.fm, como ilustrado nas etapas 2.3 (Obtenção de itens similares da Last.fm) e 2.4 (Comparação entre listas de coocorrências e listas de itens similares da Last.fm) da

metodologia. A base de dados de itens similares da Last.fm foi escolhida por ser de um grande domínio público, conhecida como uma base robusta por ter sido gerada por uma plataforma de serviços musicais consolidada e reconhecida, com um volume extenso de usuários em todo o mundo.

Para calcular a similaridade entre o conjunto de itens que coocorrem com o item j (obtidos da lista de coocorrências) e o conjunto de itens similares ao item j (obtidos da lista da Last.fm) foi desenvolvida a métrica M_j apresentada na Equação 1. Essa métrica retorna o percentual de itens que coocorrem com j e também estão presentes na lista de similaridades da Last.fm para o mesmo item j. Desse modo, a métrica é definida como a cardinalidade do conjunto resultante da interseção entre o conjunto A (itens que coocorrem com j) e o conjunto B (itens similares ao item j), dividido pela cardinalidade do conjunto A.

$$M_j = \frac{|A \cap B|}{|A|} \quad (1)$$

Por exemplo, suponha que a lista de coocorrências contenha a informação de que o artista x coocorre nas programações das rádios com os artistas w, y e z, ou seja, $A = \{w, y, z\}$. Por outro lado, a lista de artistas similares ao artista x (segundo a Last.fm) é dada pelo conjunto $B = \{y, k, t, v, m, n\}$. Sendo assim, a similaridade entre essas listas de acordo com a métrica proposta é $M_x = 1/3$, uma vez que temos apenas um artista em comum nos conjuntos A e B e três artistas contidos no conjunto A.

Além da métrica de similaridade elaborada (Equação 1), utiliza-se também o coeficiente de similaridade de Jaccard [5], também conhecido como índice Jaccard, como uma alternativa adicional para se fazer a análise de similaridade entre as listas de coocorrências e as listas de itens similares da Last.fm. O cálculo do índice de Jaccard é demonstrado na Equação 2.

$$J_j = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Seguindo o exemplo utilizado para a Equação 1, o resultado do índice de Jaccard entre A e B, para o artista x, é $J_x = 1/8$. Nesse caso, estamos levando em consideração também o tamanho da lista de artistas similares ao artista x, de acordo com a Last.fm. No entanto, os tamanhos entre as listas vindas da Last.fm e as listas de coocorrências das rádios podem apresentar tamanhos muito diferentes, não representando bem os resultados dos índices.

Portanto, além do próprio índice de Jaccard, também utiliza-se nas análises o quociente da divisão entre os índices de Jaccard obtidos e os tamanhos das listas de coocorrências correspondentes, demonstrado na Equação 3.

$$MJ_j = \frac{J_j}{|A|} \quad (3)$$

Aproveitando os exemplos dados para as Equações 1 e 2, o resultado da divisão entre J_x e A, para o artista x, é $MJ_x = (1/8)/3$.

A seguir apresenta-se a justificativa para o uso das três métricas nas análises. A primeira métrica (Equação 1) é utilizada para se saber a porcentagem dos itens das listas de coorrências (A) que estão contidos nas respectivas listas de itens similares da Last.fm (B). Já o índice de Jaccard (Equação 2) é usado para comparar a lista A com a lista B, verificando o grau de similaridade entre as duas listas. Por fim, temos a última métrica elaborada (Equação 3) na tentativa de manter mais justos os resultados de Jaccard, pois com a aplicação dos filtros as listas de coocorrências ficam cada vez menores em relação às listas de itens similares da Last.fm, impactando nos resultados obtidos por Jaccard, podendo tornar os resultados tendenciosos.

4. RESULTADOS EXPERIMENTAIS

Nesta seção são apresentadas as análises de similaridade feitas para artistas (Seção 4.1), músicas (Seção 4.2) e gêneros musicais (Seção 4.3).

4.1 Análise para artistas

Em relação aos resultados da análise para artistas, a Tabela 2 apresenta os valores médios de similaridade obtidos por uso da métrica de similaridade apresentada na Equação 1. Nessa tabela, a primeira coluna indica o tipo de filtro utilizado na geração da base de dados transacional (ver Seção 3.1). As demais colunas mostram a similaridade média entre as listas de coocorrências e as listas de itens similares da Last.fm (Sim.) e o tamanho médio das listas de coocorrências (Tam. List.) para três valores diferentes de suporte mínimo: 5%, 10% e 15%.

A Tabela 2 mostra que, para o suporte mínimo de 5% foram obtidas listas de coocorrências com tamanhos médios entre 81,5 e 186,3 itens. Já os resultados dos cálculos de similaridade entre as listas de coocorrências e de itens similares da Last.fm ficaram entre 0,23 e 0,34. À medida em que o filtro de ocorrências mínimas de artistas aumenta, o tamanho médio das listas de coocorrências diminui e o valor médio de similaridade aumenta. O tamanho médio das listas diminui porque quanto maior é o filtro, menor é o número de itens que compõem cada transação da base de dados transacional. Já o valor médio de similaridade aumenta porque as listas de coocorrências ficam mais restritas, permanecendo somente itens que estão mais correlacionados.

Tabela 2: Similaridade entre listas (Artistas)

	Suporte 5%		Suporte 10%		Suporte 15%	
Filtro	**Sim.**	**Tam. List.**	**Sim.**	**Tam. List.**	**Sim.**	**Tam. List.**
Sem filtro	0,23	186,3	0,39	49,8	0,59	11,6
3	0,27	138,3	0,47	32,0	0,72	6,0
5	0,30	117,5	0,49	24,8	0,66	3,2
10	0,34	81,5	0,55	15,3	0,83	1,5

Para o suporte mínimo de 15%, foram obtidas listas de coocorrências com tamanhos médios entre 1,5 e 11,6. Os resultados dos cálculos de similaridade entre essas listas e as listas de itens similares da Last.fm ficaram entre 0,59 e 0,83, apresentando valores médios maiores do que aqueles valores obtidos para o suporte mínimo de 5%. Ao analisar esses números percebe-se que, além do filtro de ocorrências mínimas, quando se aumenta o suporte mínimo, as listas de coocorrências também ficam mais restritas, fazendo com que seus itens tendam a estar cada vez mais correlacionados.

Os resultados apresentados na Tabela 2 mostram que com o suporte mínimo e filtro de ocorrências mínimas baixos, o tamanho das listas de coocorrências geradas pode ser grande e os itens podem não estar tão correlacionados. Por outro

lado, quanto maiores são esses filtros, menores são as listas de coocorrências e mais correlacionados são os itens nelas contidos, fazendo com que, geralmente, os valores médios de similaridade entre essas listas e as listas de itens similares da Last.fm sejam cada vez maiores.

A Tabela 3 apresenta, para a análise de artistas, os valores médios dos índices de Jaccard obtidos entre as listas de coocorrências e as listas de itens similares da Last.fm. A primeira coluna representa o tipo de filtro utilizado na geração da base de dados transacional. As demais colunas apresentam o índice de Jaccard médio entre as listas de coocorrências e as listas de itens similares da Last.fm (Jac.) e o resultado médio da divisão entre cada índice Jaccard obtido e o tamanho da lista de coocorrências correspondente (Jac. List.). Os valores de suporte utilizados foram os mesmos: 5%, 10% e 15%.

Tabela 3: Similaridade Jaccard entre listas (Artistas)

	Suporte 5%		Suporte 10%		Suporte 15%	
Filtro	**Jac.**	**Jac. List.**	**Jac.**	**Jac. List.**	**Jac.**	**Jac. List.**
Sem filtro	0,1216	0,0014	0,0871	0,0034	0,0408	0,0057
3	0,1373	0,0019	0,0797	0,0043	0,0281	0,0070
5	0,1437	0,0021	0,0697	0,0045	0,0172	0,0065
10	0,1549	0,0026	0,0585	0,0052	0,0099	0,0083

Observa-se na Tabela 3 que, para o suporte de 5%, à medida em que o filtro de ocorrências mínimas de artistas aumenta, o índice médio de Jaccard aumenta também, juntamente com o quociente médio entre Jaccard e o tamanho da lista de coocorrências. Porém, quando se aumenta o suporte (10% e 15%), o índice médio de Jaccard diminui, ao passo que o quociente entre Jaccard e o tamanho da lista aumenta.

O índice de Jaccard diminui pois quanto maiores os valores de suporte e os filtros aplicados, menores ficam as listas de coocorrências, se distanciando do tamanho das listas de itens similares da Last.fm. Por outro lado, quando se observa o quociente médio da divisão do Jaccard pelo tamanho da lista de coocorrências, percebe-se que o mesmo aumenta à medida em que os filtros aplicados ficam mais rigorosos, conforme também observado pela métrica utilizada na Tabela 2.

A Tabela 4 apresenta algumas novas correlações encontradas entre artistas que coocorrem nas programações das rádios e que não são considerados similares pela Last.fm. As configurações utilizadas para essa análise foram as seguintes: suporte mínimo do Apriori de 10% e filtro de ocorrências mínimas de 10 itens. Embora não sejam considerados similares pela Last.fm, observou-se que esses artistas compartilham características, como por exemplo o estilo musical, representando conhecimento novo descoberto através da metodologia apresentada neste artigo, e que não foi descoberta pela Last.fm.

Tabela 4: Novas correlações encontradas entre artistas, que não estavam presentes na Last.fm

Artista	Novas Correlações
Alan Jackson	Carrie Underwood
Rihanna	Enrique Iglesias, Jason Mraz, Coldplay
Beyonce	Taio Cruz, Black Eyed Peas, Maroon 5
The Rolling Stones	Steve Miller Band, Tom Petty
3 Doors Down	Red Hot Chili Peppers, U2, Green Day
Green Day	Kings Of Leon, Train, U2

4.2 Análise para músicas

A Tabela 5 apresenta a mesma estrutura da Tabela 2, porém os dados são relativos à análise de similaridade realizada para músicas. Para o suporte mínimo de 5% foram obtidas listas de coocorrências com tamanho médio entre 58,6 e 225,5. As similaridades médias entre essas listas e as listas de itens similares da Last.fm ficaram entre 0,20 e 0,35. Esses resultados seguem o mesmo padrão de comportamento daqueles apresentados nas análises para artistas, ou seja, à medida em que o filtro de ocorrências mínimas de músicas aumenta, o tamanho médio das listas de coocorrências diminui e, geralmente, o valor médio de similaridade entre essas listas e as listas de itens similares da Last.fm aumenta.

Tabela 5: Similaridade entre listas (Músicas)

	Suporte 5%		Suporte 10%	
Filtro	**Sim.**	**Tam. List.**	**Sim.**	**Tam. List.**
Sem filtro	0,20	225,5	0,42	12,9
3	0,25	149,2	0,47	6,2
5	0,29	110,8	0,45	5
10	0,35	58,6	0,60	3

Para o suporte mínimo de 10%, foram obtidas listas de coocorrências com tamanho médio entre 3 e 12,9. Os resultados de similaridade média entre essas listas e as listas de itens similares da Last.fm ficaram entre 0,42 e 0,60, apresentando assim valores médios maiores do que aqueles obtidos para o suporte mínimo de 5%. Assim como para as bases de dados de artistas, quando se aumenta o suporte mínimo, as listas de coocorrências ficam mais restritas, fazendo com que seus itens tendam a estar cada vez mais correlacionados.

A Tabela 6 segue o mesmo formato da Tabela 3, porém apresenta resultados relativos à analise para músicas. Essa tabela demonstra que, tanto para o suporte de 5% quanto para o suporte de 10%, à medida em que o filtro de ocorrências mínimas de artistas aumenta, o índice médio de Jaccard diminui e o quociente entre Jaccard e o tamanho da lista aumenta.

Tabela 6: Similaridade Jaccard entre listas (Músicas)

	Suporte 5%		Suporte 10%	
Filtro	**Jac.**	**Jac. List.**	**Jac.**	**Jac. List.**
Sem filtro	0,0724	0,0005	0,0135	0,0016
3	0,0640	0,0007	0,0088	0,0018
5	0,0568	0,0009	0,0085	0,0017
10	0,0439	0,0012	0,0050	0,0023

4.3 Análise para gêneros musicais

A Tabela 7 apresenta a mesma estrutura das Tabelas 2 e 5, porém os dados são relativos à análise de similaridade para gêneros musicais. Observa-se que para o suporte mínimo de 5% foram obtidas listas de coocorrências com tamanhos médios entre 32 e 51,1. As similaridades médias entre essas listas e as listas de itens similares da Last.fm ficaram entre 0,15 e 0,21. Nota-se que esses resultados seguem o mesmo padrão de comportamento dos resultados apresentados nas análises para artistas e músicas, em que ao se aumentar o filtro de ocorrências mínimas de itens, o tamanho médio das

listas de coocorrências diminui e, geralmente, a similaridade média entre essas listas e a da Last.fm aumenta.

Tabela 7: Similaridade entre listas (Gêneros)

	Suporte 5%		Suporte 15%		Suporte 20%	
Filtro	Sim.	Tam. List.	Sim.	Tam. List.	Sim.	Tam. List.
Sem filtro	0,15	51,1	0,26	21,6	0,38	15,3
3	0,18	40,5	0,43	17,1	0,58	9,9
5	0,21	35,9	0,45	13,7	0,56	7,9
10	0,18	32	0,54	10,5	0,66	5,1

Para o suporte mínimo de 20%, foram obtidas listas de coocorrências com tamanhos médios entre 5,1 e 15,3. Os resultados de similaridade média entre essas listas e as listas da Last.fm ficaram entre 0,38 e 0,66, apresentando assim valores médios maiores do que aqueles obtidos para o suporte mínimo de 5%. Assim como para as bases de dados de artistas e de músicas, quando se aumenta o suporte mínimo, as listas de coocorrências ficam mais restritas, fazendo com que seus itens tendam a estar cada vez mais correlacionados.

A Tabela 8 segue o mesmo formato das Tabelas 3 e 6, porém apresenta resultados relativos à analise para gêneros musicais. A Tabela 8 mostra que quando o suporte aumenta, até certo ponto (de 5% a 15%), o índice de Jaccard também aumenta. Mas a partir do suporte de 15% com filtro de 5, o índice de Jaccard médio começa a cair. Já o quociente médio entre Jaccard e o tamanho da lista aumenta.

Tabela 8: Similaridade Jaccard entre listas (Gêneros)

	Suporte 5%		Suporte 15%		Suporte 20%	
Filtro	Jac.	Jac. List.	Jac.	Jac. List.	Jac.	Jac. List.
Sem filtro	0,0702	0,0018	0,0918	0,0040	0,0973	0,0066
3	0,0730	0,0025	0,0980	0,0076	0,0754	0,0110
5	0,0758	0,0031	0,0890	0,0080	0,0659	0,0108
10	0,0844	0,0023	0,0891	0,0101	0,0567	0,0131

Os resultados das análises para gêneros musicais apresentam características semelhantes às análises de resultados para artistas e para músicas, variando-se apenas os valores apresentados.

Tabela 9: Novas correlações encontradas entre gêneros musicais, que não estavam presentes na Last.fm

Gênero	Novas Correlações
Thrash Metal	Rock
Glam Rock	Pop, New Wave, Progressive Rock
Rap	Pop, Soul
Heavy Metal	Rock, New Wave, Progressive Rock
Grunge	Classic Rock, New Wave, Punk Rock
House	Rock, Pop, Soul, Hip-Hop

A Tabela 9 segue a mesma estrutura da Tabela 4 (Seção 4.1), apresentando algumas novas correlações encontradas entre gêneros musicais que coocorrem nas programações das rádios e que não são considerados similares pela Last.fm. As configurações utilizadas para essa análise foram as seguintes: suporte mínimo do Apriori de 10% e filtro de ocorrências mínimas de 10 itens. Como exemplos, temos coocorrências como Thrash Metal e Rock e também Heavy Metal e Rock, mostrando assim que a metodologia proposta é capaz de descobrir conhecimento novo a respeito da correlação entre gêneros musicais.

5. LIMITAÇÕES DO ESTUDO

Nesta seção são discutidas as limitações das análises realizadas, baseando-se nas categorias de ameaças à validade apresentadas por [14]. A seguir, são listadas as ameaças levantadas e as medidas tomadas para reduzi-las.

Validade interna: Uma ameaça à validade interna do estudo é a escolha do algoritmo para extração dos conjuntos de itens frequentes e o suporte mínimo utilizado. Para tentar minimizar esse risco, utilizou-se o Apriori que é um algoritmo clássico para extração de conjuntos de itens frequentes. Além disso, foram utilizados variados valores de suporte para minimizar o risco da escolha de apenas um valor de suporte, que poderia não ser o adequado.

Validade externa: Uma das ameaças identificadas para a validade externa foi a escolha correta das rádios a serem utilizadas para as coletas de dados de programações musicais e do *website* a ser utilizado para a obtenção dos conjuntos de itens musicais similares (Last.fm). Além disso, garantir que esses dados coletados representem a realidade das programações das rádios e também das correlações entre os itens musicais.

Em relação às rádios utilizadas no estudo (Seção 3.1), as coletas dos dados foram realizadas de forma automática por serviços de terceiros. Sendo assim, pode acontecer de determinadas rádios terem suas programações monitoradas mais regularmente do que outras. Na tentativa de se extrair os conjuntos de dados de rádios que melhor refletissem a realidade de suas programações, selecionou-se de maneira automática os dados de programações das rádios que foram coletados mais regularmente.

Já o *website* Last.fm, escolhido como fonte para obter os conjuntos de itens musicais similares, como qualquer outra fonte pública ou privada que pudesse ser aplicada, não garante que esses itens sejam realmente correlacionados. Porém, esse serviço foi escolhido por ser uma das principais plataformas de recomendação de músicas no mundo, contando com a colaboração de milhões de usuários para as classificações dos itens musicais, gerando uma rica fonte de conhecimento coletivo, principalmente no que tange à similaridade entre artistas.

Validade de construção: identificou-se como uma ameaça de construção o cenário em que itens (músicas, artistas e gêneros musicais) iguais, provenientes de fontes distintas (rádios e Last.fm), fossem considerados diferentes por apresentarem alterações no título de uma fonte para a outra. Para tentar resolver esse tipo de problema, foi aplicada a distância de edição, através do algoritmo de Levenshtein [10], para a comparação dos títulos dos itens.

Validade de conclusão: uma ameaça à validade de conclusão do estudo é em relação à métrica utilizada para avaliar a similaridade entre os conjuntos de itens frequentes, extraídos das rádios, e os conjuntos de itens similares, extraídos da Last.fm. Embora tenha sido elaborada por este trabalho, não sendo uma métrica solidamente utilizada por demais estudos, foi a maneira mais coerente encontrada para se extrair uma taxa que representasse o grau de interseção entre os conjuntos de itens, para o problema em análise. Esta métrica é baseada no coeficiente de similaridade de Jaccard.

6. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foram utilizados metadados de programações de estações de rádio com intuito de investigar se os itens (artistas, músicas e gêneros musicais) que coocorrem nessas programações podem ser considerados similares entre si, de acordo com o conceito de similaridade adotado pela Last.fm. As listas de coocorrência de itens que foram comparadas com as listas de itens similares retornadas pela Last.fm, foram construídas aplicando-se o algoritmo Apriori em bases de dados transacionais geradas a partir de metadados de programações de rádios.

Com base nos resultados experimentais obtidos, observou-se que é possível encontrar correlação entre os itens que coocorrem nas programações das rádios. Também verificou-se que, quanto mais restrito é o conjunto de itens analisado (aumentando-se o filtro de ocorrências mínimas de itens e o suporte do Apriori), maior é a similaridade entre as listas de coocorrências de itens das programações das rádios e as listas de itens similares produzida pela Last.fm. No entanto, novas correlações entre itens foram descobertas, ou seja, foram encontrados itens que coocorrem nas programações das rádios de forma muito frequente, mas que não são considerados similares pela Last.fm.

As descobertas realizadas são úteis tanto para evidenciar a existência da correlação entre os itens musicais das programações das rádios, quanto para melhorar o desempenho de sistemas de recomendação como a própria Last.fm, que tipicamente fazem uso de dados de similaridade para aumentar a qualidade da recomendação. Portanto, conclui-se que a metodologia proposta não somente é capaz de encontrar itens realmente similares a partir da comparação com a lista de similaridade da Last.fm, como também é capaz de identificar novos itens similares que não haviam sido identificados pela Last.fm. De forma geral, conclui-se também que os dados de programação de rádios são bastante ricos para a descoberta de conhecimento sobre correlações entre itens.

Como trabalho futuro, almeja-se realizar análises em bases de metadados de programações de rádios visando identificar quais itens aumentam a probabilidade de ocorrência de outros itens. Outra sugestão de trabalho futuro é estender o trabalho para um maior número de rádios, visando descobrir ainda mais similaridades entre itens.

7. AGRADECIMENTOS

Os autores agradecem às instituições UFOP, CAPES, FAPEMIG e CNPq por apoiarem o desenvolvimento desta pesquisa e ao projeto Radialize pelo fornecimento dos dados para as análises.

8. REFERÊNCIAS

[1] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, volume 1215, pages 487–499, 1994.

[2] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He. Music recommendation by unified hypergraph: Combining social media information and music content. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 391–400, New York, NY, USA, 2010. ACM.

[3] G. Geleijnse, M. Schedl, and P. Knees. The quest for ground truth in musical artist tagging in the social web era. In *International Society for Music Information Retrieval (ISMIR)*, pages 525–530. Citeseer, 2007.

[4] J. Hong, H. Deng, and Q. Yan. Tag-based artist similarity and genre classification. In *IEEE International Symposium on Knowledge Acquisition and Modeling Workshop, 2008. KAM Workshop 2008.*, pages 628–631. IEEE, 2008.

[5] P. Jaccard. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge, 1901.

[6] N. Koenigstein, G. Dror, and Y. Koren. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. In *Proceedings of the 50th ACM Conference on Recommender Systems*, pages 165–172. ACM, 2011.

[7] Last.fm. Last.fm web services: artist.gettoptags. `http://www.last.fm/api/show/artist.getTopTags`, 2016. [acesso em 10/02/2016].

[8] Y. Liu, Y. Wang, A. Shenoy, W.-H. Tsai, and L. Cai. Clustering music recordings by their keys. In *International Society for Music Information Retrieval (ISMIR)*, pages 319–324, 2008.

[9] S. Mahajan, A. Reshamwala, N. Sharma, D. Vineet, A. Sharma, and P. Shah. Prediction of yahoo! music sequences on user's musical taste. In *Proceedings of the International Conference on Advances in Information Technology*, pages 6–9. AIT, 2012.

[10] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, 33(1):31–88, 2001.

[11] K. Neubarth, I. Goienetxea, C. Johnson, and D. Conklin. Association mining of folk music genres and toponyms. In *International Society for Music Information Retrieval (ISMIR)*, pages 7–12, 2012.

[12] F. Pachet, G. Westermann, and D. Laigre. Musical data mining for electronic music distribution. In *Proceedings of the 1th International Conference on Web Delivering of Music, 2001.*, pages 101–106. IEEE, 2001.

[13] W. N. Venables, D. M. Smith, and R. D. C. Team. An introduction to R, 2002.

[14] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[15] Y. Ye and C.-C. Chiang. A parallel apriori algorithm for frequent itemsets mining. In *Fourth International Conference on Software Engineering Research, Management and Applications, 2006.*, pages 87–94. IEEE, 2006.

[16] A. Ziesemer and J. Oliveira. How to know what do you want? a survey of recommender systems and the next generation. In *Proceedings of the Eighth Brazilian Symposium on Collaborative Systems, SBSC*, pages 104–111, 2011.