Mapeando o universo da mídia usando dados gerados por usuários em redes sociais online.

Pedro H. F. Holanda[1], **Bruno Guilherme**[1], **João Paulo V. Cardoso**[2]
Ana Paula Couto da Silva[1], **Olga Goussevskaia**[1]

[1]Departamento de Ciência da Computação - UFMG
[2]Departamento de Engenharia Elétrica - UFMG

{holanda,brunoguilherme,ana.coutosilva,olga}@dcc.ufmg.br
jpcardoso@ufmg.br

Abstract. *The way people watch movies and TV has been going through great changes in recent years, one being that people are increasingly willing to share their TV watching habits with friends (and strangers) through online social network platforms. In this work, we collect usage data from the online social network tvtag and propose a data structure to represent and efficiently retrieve similarity information about movies and TV shows. We refer to this structure as "map of media", because it consists of a multi-dimensional Euclidean space, where each item is represented by a set of coordinates and the distance between them represents the similarity. We propose several metrics to evaluate the resulting structure and show that, besides computational efficiency, the proposed approach provides a high quality measure for item similarity. Moreover, the quality increases with increasing dimensionality of the map.*

Resumo. *A maneira como as pessoas consomem diferentes mídias está sofrendo grandes mudanças nos últimos anos. Uma das mais interessantes é que as pessoas desejam compartilhar seus hábitos e preferências através das redes sociais online. Neste trabalho, a partir dos dados gerados por usuários na rede social tvtag, nós propomos uma estrutura de dados para representar e consultar de forma eficiente informação de similaridade entre filmes e programas de TV. Nos referimos a esta estrutura de dados como "mapa da mídia", pois a mesma consiste de um espaço Euclidiano multidimensional, onde cada item é representado por um conjunto de coordenadas e a distância entre dois itens representa a similaridade. Nós propomos várias métricas para avaliar a estrutura resultante e mostramos que, além da eficiência computacional, a mesma provê uma medida de similaridade com alta qualidade. Além disso, a qualidade tende a aumentar com o aumento da dimensionalidade do mapa.*

1. Introdução

A maneira como as pessoas assistem a programas de TV e a filmes tem mudado drasticamente nos últimos anos [Torrez-Riley 2011]. Atualmente, as pessoas assistem aos seus programas preferidos através da internet com conexões de alta velocidade e utilizando várias fontes diferentes (Youtube e Netflix[1]). Além dos computadores, as pessoas assistem a estes conteúdos a partir de smartphones, tablets e smart TVs

[1]www.youtube.com, www.netflix.com

[Raje 2014, Sareen 2014]. Um outro ponto interessante é que a quantidade e diversidade de conteúdo disponível aumenta cada vez mais à medida que as tecnologias disponíveis para a produção de vídeo se tornam mais acessíveis.

Entre os diversos setores de produção de mídias, podemos destacar a televisão e o cinema. Assistir programas de TV é um hábito social, que gera discussões em diferentes ambientes do nosso dia-a-dia. Da mesma forma que a maneira de assistir aos programas mudou nos últimos anos, o modo de discutir e compartilhar opiniões em torno do conteúdo televisivo (ou de cinema) também sofreu modificações: as pessoas passaram a utilizar as redes sociais online como meio de propagação de opiniões e preferências [Narasimhan and Vasudevan 2012, Torrez-Riley 2011]. Para tal, são utilizadas redes sociais genéricas, como o Twitter ou redes sociais específicas, como o tvtag[2].

Apesar das mudanças nos hábitos de consumo e compartilhamento de opiniões do conteúdo da TV e do cinema, a maneira como as pessoas navegam pelas coleções existentes de programas e descobrem conteúdos novos não mudou muito. Na maioria das vezes, a navegação é realizada através de listas sequenciais, em ordens alfabéticas ou hierárquicas, ou ordenadas pela programação dos canais e guias de TV.

Poucos trabalhos na literatura propõem estratégias mais elaboradas para a navegação através de conteúdos de diferentes mídias. Podemos destacar algumas abordagens voltadas para o domínio da música [Knees et al. 2006, Neumayer et al. 2005, Goussevskaia et al. 2008, Watchmi 2014]. Os autores em [Goussevskaia et al. 2008] apresentam uma abordagem de navegação em coleções de músicas baseada no assim chamado "mapa da música", que consiste de um espaço Euclidiano multidimensional, onde cada item é representado por um conjunto de coordenadas e a distância entre dois itens representa a similaridade. Neste trabalho, estendemos esta abordagem para um novo domínio: cinema e televisão. Além disso, aprimoramos a metodologia e propomos novas métricas para avaliar a estrutura resultante no que tange o agrupamento de conteúdos similares (por exemplo, em gênero e tipo).

Para construir e analisar um sistema de organização e navegação em conteúdo baseado em similaridade, duas questões importantes devem ser respondidas: (1) Como definir similaridade entre pares de itens? Note que esta é uma questão essencialmente subjetiva, já que, o que pode parecer similar para uma pessoa, pode parecer discrepante para outra, dependendo do seu grau de expertise no assunto, por exemplo; (2) Qual estrutura de dados devemos usar para representar e consultar de forma eficiente as informações de similaridade de conteúdos, considerando robustez e adequação para navegação em grandes quantidades de conteúdo?

Neste trabalho, partimos da hipótese de que similaridade pode ser derivada a partir de dados gerados pelos usuários em uma rede social online voltada para fãs de cinema e TV, em particular o tvtag. Disponibilizada em 2010, a popularidade do tvtag teve um aumento notório: o número de usuários cresceu de 30.000 para 4.5M em 2013, com aproximadamente 500M de atividades realizadas pelos usuários[3]. Nesta rede social, os usuários podem fazer *check-ins* nos programas de TV, publicar avaliações (através de *likes* e *dislikes*) e trocar opiniões sobre um determinado conteúdo.

[2] `www.twitter.com`, `www.tvtag.com`

[3] `blog.tvtag.com`

Para a estrutura de dados de navegação, propomos o "mapa da mídia", estendendo a abordagem apresentada em [Goussevskaia et al. 2008]. Com isso, similaridade de n itens pode ser representada com complexidade espacial $O(n)$, e a similaridade entre dois itens pode ser computada com complexidade temporal constante, ou seja, $O(1)$. A partir desta estrutura, organizamos os conteúdos segundo o grau de similaridade entre os mesmos, gerando uma navegação mais aprimorada entre coleções de programas de TV e filmes, baseada em preferências dos usuários.

As contribuições deste trabalho podem ser resumidas da seguinte forma: (1) Coletamos dados sobre atividades de usuários do tvtag e fizemos uma caracterização inicial dos mesmos. Em particular, mostramos que a distribuição das atividades entre as diversas mídias segue uma lei de potência; que os usuários do tvtag tendem a ser mais positivos em relação ao conteúdo disponibilizado e tem como preferência os gêneros de comédia e animação (Seção 2); (2) Através da análise de co-ocorrência de itens em históricos de atividade de usuários, nós propusemos uma medida de similaridade para pares de itens (Seção 3); (3) A partir da medida par-a-par de similaridade, construímos uma estrutura de dados que chamamos de "mapa da mídia" (Seção 3); (4) Propusemos métricas automatizadas para avaliar a qualidade da medida de similaridade obtida e pudemos observar que a qualidade da medida aumenta com maior dimensionalidade, que dados similares são agrupados por regiões e que transições de similaridade entre itens consecutivos ao longo de trajetórias aleatórias no mapa são suaves (Seção 4); (5) Finalmente, realizamos uma análise manual dos mapas obtidos em várias dimensões, recorrendo a opiniões (subjetivas) de pessoas "comuns" e de "experts" no assunto de cinema e televisão (Seção 4). Concluímos o artigo com uma discussão de trabalhos relacionados (Seção 5) e considerações finais (Seção 6).

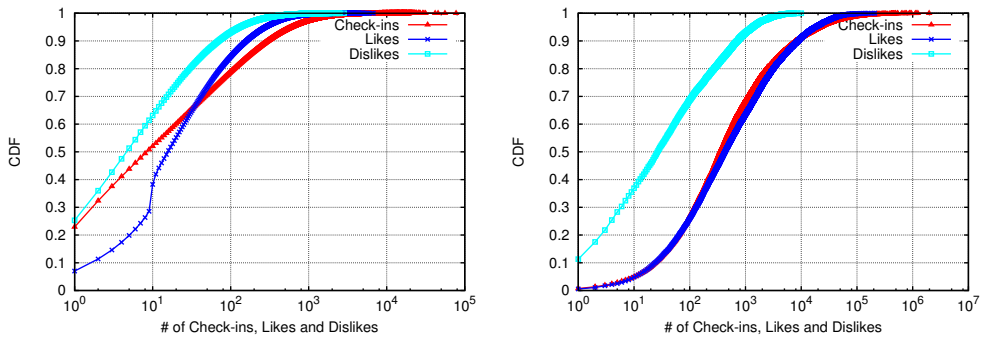
2. Coleta de Dados e Caracterização

Este trabalho é pioneiro na coleta e análise da rede social online tvtag. Assim, esta seção tem como objetivo apresentar alguns resultados importantes sobre o comportamento dos usuários e da distribuição dos gêneros dos programas de TV e filmes encontrados na nossa base de dados.

Base de dados: Para definir o mapa da mídia, duas fontes diferentes de dados foram usadas: a rede social online tvtag e o TMDB (The Open Movie Database[4]). A partir do tvtag foram coletadas informações sobre preferências e atividades dos usuários relacionadas ao conteúdo de TV e filmes, e a partir do TMBD foram obtidos metadados, tais como data de estreia, atores, diretores e gêneros dos programas de TV e filmes. A classificação de gêneros a partir do TMDB será essencial para a caracterização do conteúdo do tvtag e para a validação da estrutura de dados baseada em similaridade.

Coleta de dados: Para a coleta de dados, foram implementados dois *crawlers* Web. O primeiro *crawler* coletou as atividades dos usuários (check-ins, likes e dislikes) no tvtag. Nosso conjunto de dados representa uma boa cobertura da rede do tvtag no período entre 2011 e 2012, e consiste no total de 29M check-ins, 21M likes e 1M dislikes. O segundo *crawler* coletou os metadados disponibilizados pelo TMDB. Foram coletados 100% dos dados de programas de TV e filmes disponíveis no TMDB. Após o cruzamento dos dois conjuntos de dados, foi possível obter tanto a atividade dos usuários quanto os

[4] www.themoviedb.org



(a) CDF por usuário (excluindo os usuários com 0 check-ins (1M), likes (900K) e dislikes (1,6M)).

(b) CDF por programa de TV (excluindo os programas com 0 check-ins (76), likes (6) e dislikes (910)).

Figura 1. Distribuição de atividades (check-ins, likes e dislikes) no tvtag.

Usuários	1.745.000
Filmes	9.300
Programas de TV	5.000
Gêneros (TMDB)	26
Itens c/ gênero	9.977
Check-ins	92.077.000
Likes	52.776.000
Dislikes	3.033.000

Programa de TV	Likes	Check-ins	Dislikes
Big Bang Theory	206.769	1.972.968	4.935
Family Guy	206.458	459.463	6.428
Simpsons	192.276	378.779	4.449
House	180.631	405.951	3.380
Walking Dead	159.704	159.322	2.444
Glee	153.192	949.049	8.403
HIMYM	143.998	537.257	4.095
True Blood	143.888	1,238.715	5.253
South Park	139.220	153.792	5.216
Dexter	139.011	749.978	3.598

Tabela 1: (a) Dataset tvtag e TMDB (b) Top 10 programas de TV no tvtag.

metadados de aproximadamente 3.000 programas de TV e 7.000 filmes. As estatísticas relacionadas aos dados coletados estão descritas na Tabela 1(a). Por limitações de espaço, a caracterização apresentada a seguir foca nos programas de TV encontrados nos dados coletados.

Caracterização das atividades dos usuários: As Figuras 1(a) e 1(b) mostram a CDF (Função de Distribuição Acumulada) do total de check-ins, likes e dislikes realizados por cada usuário e em cada programa de TV. Para facilitar a visualização dos resultados, foram removidos dos gráficos os usuários e programas com zero check-ins, likes e dislikes. Podemos observar que quase 52% dos usuários fizeram até 10 check-ins. Aproximadamente 50% dos programas receberam até 350 check-ins. 20% dos usuários são altamente engajados na rede social, realizando entre 100 e 1.000 check-ins no período coletado e 3% dos usuários realizaram mais de 1.000 check-ins. Para termos certeza de que estes usuários não são robôs, avaliamos manualmente o comportamento dos mesmos. Segundo nossas análises, eles possuem comportamento similar aos dos usuários que interagem com outros usuários e programas de TV. Considerando o nosso conjunto de dados, uma fração considerável de programas de TV (9%) recebe mais de 10.000 check-ins. Estes valores podem ser usados como uma estimativa de audiência destes programas.

As Figuras 1(a) e 1(b) também mostram como as pessoas utilizam os

likes/dislikes. Os usuários do tvtag tendem a ser mais positivos do que negativos em relação aos programas de TV. A distribuição do número de dislikes tem um decaimento exponencial mais acentuado quando comparada as distribuições de check-in e like. Nos dados coletados, temos aproximadamente 52M likes e 3M de dislikes. Mais de 60% dos usuários atribuem mais de 10 likes. No entanto, somente 35% dos usuários atribuem o mesmo número de dislikes. A Tabela 1(b) apresenta a lista dos top 10 programas de TV considerando o total de likes.

Distribuição de Gêneros: A partir da informação coletada do TMDB, categorizamos aproximadamente 3.000 programas de TV, cadastrados no tvtag, considerando os 26 gêneros diferentes (TMDB). Nos dados coletados, 63% dos programas são classificados com um gênero; 24% com dois gêneros e 13% com três gêneros ou mais. Além disso, Drama é o gênero mais predominante no TMDB, enquanto Comédia é o favorito no tvtag. Esta diferença de porcentagem de conteúdo para diferentes gêneros pode ser explicada pelas diferentes faixas etárias da audiência. Redes sociais online tendem a atrair a audiência mais jovem quando comparado com o público da TV em geral.

3. Navegação Baseada em Similaridade

O processo para obter a estrutura de dados "mapa da mídia" a partir da informação coletada do tvtag é composto por três etapas: primeiro, estimamos similaridade item-a-item utilizando técnicas de filtragem colaborativa; segundo, construímos um grafo a partir desses valores de similaridades item-a-item; terceiro, mapeamos o grafo em um espaço Euclidiano, preservando aproximadamente as distâncias.

3.1. Similaridade Item-a-Item

Para obter valores de similaridade entre os filmes e programas de TV, nós usamos técnicas baseadas em filtragem colaborativa. O fato de dois itens serem relacionados porque co-ocorrem frequentemente em dados de uso demonstrou funcionar bem em estudos anteriores [Linden et al. 2003, Goussevskaia et al. 2008, Kuhn and Wattenhofer 2007], embora possua uma limitação em relação à amostra de usuários utilizada. Assim como a *Amazon* usa o fato de dois itens estarem relacionados por terem sido comprados pela mesma pessoa, presumimos que duas séries ou dois filmes são relacionadas caso recebam *like* do mesmo usuário, revelando suas preferências.

A simples contagem do número de ocorrência de dois itens para calcular a similaridade pareada superestima a similaridade de itens populares, uma vez que estes claramente possuem maior probabilidade de aparecerem no histórico de *likes* de um mesmo usuário devido ao seu alto número de ocorrências. Para superar esse problema, algum tipo de normalização se faz necessária. Vários coeficientes foram propostos para abordar essa questão [Matsuo et al. 2007], e o coeficiente utilizado em nossas análises foi o coeficiente cosseno, definido por $cos(i,j) = n_{i,j}/\sqrt{n_i n_j}$, onde $n_{i,j}$ é o número de co-ocorrências dos itens i e j, e n_i (n_j) o número de ocorrências individuais do item i (j).

3.2. Grafo de Similaridades

O processo descrito na seção anterior fornece um conjunto de medidas de similaridade item-a-item para aqueles itens que tiveram co-ocorrência significativa no tvtag. Note

que esse tipo de dado costuma criar matrizes esparsas, ou seja, somente uma pequena fração de todos os possíveis pares de itens terão um valor de similaridade conhecido. Para computar os valores de similaridades faltantes, nós construímos um grafo (não completo) de similaridades, onde cada vértice representa um filme ou um programa de TV, e arestas ponderadas são inseridas no grafo entre os vértices que possuem um valor conhecido de cosseno. Dessa forma, é possível computar a similaridade entre dois itens ao somar os pesos das arestas pertencentes ao caminho mínimo entre os mesmos no grafo.

O coeficiente cosseno resulta em valores próximos de zero para itens que não co-ocorrem frequentemente, e em valores próximos de um para itens que co-ocorrem muito frequentemente, por ser diretamente proporcional ao número de co-ocorrências. Entretanto, para traduzir a similaridade em uma medida de distância, o comportamento contrário seria necessário: itens mais similares ficam perto, e itens mais diferentes deveriam estar longe. Portanto, aplicamos o complemento da medida cosseno ($1 - cos(i, j)$) a todos os pares de itens coletados. Isso resulta em um grafo que contem uma aresta entre quaisquer dois itens que tenham aparecido juntos no histórico de likes de um mesmo usuário. Para evitar efeitos aleatórios, arestas com co-ocorrências abaixo de 2 foram removidas do grafo. Esse procedimento também elimina qualquer item que tenha ocorrido apenas uma vez. Em seguida, computamos a maior componente conexa do grafo, o que resultou em um (sub)grafo conexo G com 14.144 nós e $57.931.503$ arestas.

3.3. Mapa da Mídia

Calcular a similaridade entre dois itens baseando-se em um grafo de tamanho muito grande demanda um custoso cálculo de menor caminho se os vértices correspondentes não estiverem na vizinhança um do outro[5]. Esta avaliação de menor caminho não só implica em longos tempos de cálculo, mas também apresenta um consumo de memória extremamente elevado, já que o grafo deve estar presente inteiramente na memória, mesmo que se queira trabalhar com um subconjunto pequeno de itens.

Para o uso eficiente de um grafo tão grande em aplicativos (possivelmente móveis ou distribuídos), damos um passo além e criamos a estrutura de dados "mapa da mídia", que é um *mergulho*[6] do grafo em um espaço Euclidiano. Um mergulho é a atribuição de coordenadas para cada nó do grafo. Em nosso caso, a meta é preservar, aproximadamente, todas as distâncias par-a-par. Isto é, uma atribuição de coordenadas deve ser feita de forma que a razão $d_G(i, j)/d_E(i, j)$ entre a distância d_G no grafo e a distância d_E no mergulho seja aproximadamente igual a 1 para todos os pares de nós $(i, j) \in G$.

Usando a estrutura de dados "mapa da mídia", a distância (Euclidiana) entre os itens pode ser computada diretamente de suas coordenadas, ou seja, em tempo $O(1)$ e com $O(n)$ consumo de memória, onde n é o tamanho da coleção local do usuário. Nenhuma informação sobre quaisquer outros itens ou estruturas é necessária. Mergulhos são, portanto, particularmente adequados para aplicativos distribuídos e móveis. Além disso, um mergulho exibe várias vantagens funcionais, como noção de direção ou a possibilidade de medir volumes.

O processo de mergulho utiliza o procedimento de redução de dimensionalidade,

[5]Note que tipicamente itens não são vizinhos imediatos, já que o grafo de similaridades deve ser esparso. Um grafo denso, com $\Theta(n^2)$ arestas, seria potencialmente grande demais para ser armazenado.

[6]*Embedding*, em inglês

e existem diversas técnicas na literatura que realizam esta tarefa, tais como *Principal Component Analysis* (PCA) e *Multidimensional Scaling* (MDS) [Harel and Koren 2002, de Silva and Tenenbaum 2003]. Note que, o custo computacional dessas técnicas pode ser elevado e, ao se realizar o mergulho, a estrutura original dos dados pode ser perdida [Tenenbaum et al. 2000]. Para realizar o mergulho do grafo de similaridade em um espaço Euclidiano de dimensões reduzidas, nos inspiramos no algoritmo proposto em [Tenenbaum et al. 2000], denominado Isomap, que consiste em calcular a matriz de distâncias de todos os n^2 nós do grafo (conexo), usando, por exemplo, o algoritmo de Dijkstra e, em seguida, aplicar o algoritmo MDS clássico [Kruskal and Wish 1978] para reduzir a dimensionalidade do espaço Euclidiano gerado. Note que, apesar da elevada complexidade computacional desta técnica, foi possível, em algumas horas de computação (aproximadamente 6), em um servidor com 6 cores e 30GB de memória RAM, gerar o mergulho do nosso grafo G com 14.144 nós e $57.931.503$ arestas.

4. Avaliação dos Resultados

Para avaliar a qualidade da medida de similaridade representada pelo "mapa da mídia", utilizamos metadados obtidos de uma fonte independente, o TMDB. Como foi dito na Seção 2, conseguimos categorizar aproximadamente 3.000 programas de TV com 26 gêneros diferentes. Como a maioria dos itens pertencem a mais de um gênero, não seria suficiente simplesmente verificar se dois itens pertencem ao mesmo gênero. Primeiramente, definimos uma medida de similaridade de gêneros, utilizando a mesma metodologia que utilizamos para definir a similaridade entre conteúdos: dois gêneros são mais(menos) similares proporcionalmente a co-ocorrência dos mesmos em diferentes conjuntos de likes dos usuários. A similaridade entre dois gêneros g_i e g_j é dada por:

$$simGeneros(g_i, g_j) = cos(g_i, g_j) = \frac{cooc(g_i, g_j)}{\sqrt{itens(g_i)itens(g_j)}}, \quad (1)$$

onde $cooc(g_i, g_j)$ é o número de itens que foram classificados com ambos os gêneros, e $itens(g_i)$ e $itens(g_j)$ o número de itens que foram classificados pelo menos com gênero g_i e pelo menos com gênero g_j, respectivamente.

As métricas automáticas que iremos utilizar para medir a qualidade do mergulho são baseadas na premissa de que conteúdos de gêneros similares deveriam estar próximos após o mergulho no espaço Euclidiano.

Como, na nossa base de dados, a maioria dos itens é classificada com mais de um gênero simultaneamente, para definir quão similares são os gêneros de dois itens A e B, computamos a média dos cossenos entre os gêneros do item A e item B. Sejam as listas de gêneros atribuídos aos conteúdos A e B, $G_A =< g_1, g_2, \cdots g_k >$ e $G_B =< g_1, g_2, \cdots g_l >$, definimos a *similaridade por gênero* entre itens A e B como:

$$simPorGenero(A, B) = \frac{\sum_{g_i \in GA, g_j \in GB} simGeneros(g_i, g_j)}{kl}. \quad (2)$$

A seguir definimos as métricas que utilizam a definição acima.

Análise do gradiente de gêneros no espaço: Para o cálculo dessa métrica, a seguinte metodologia foi aplicada. Consideramos $p = 20$ pontos mais próximos a uma

reta, traçada entre a origem do mergulho e um ponto escolhido aleatoriamente no espaço Euclidiano (em d dimensões). A partir dos pontos escolhidos, ordenamos os mesmos pela distância ao ponto mais distante da origem da reta. Com a projeção desses pontos na reta traçada, percorremos a lista de gêneros atribuída a cada par consecutivo de pontos (conteúdos). O objetivo é verificar o quão suave é a transição entre gêneros dos conteúdos projetados na reta. Ou seja, iremos definir o gradiente de variação dos gêneros na reta, calculando a similaridade por gênero $simPorGenero(A, B)$, definida em (2), entre pares de pontos consecutivos dessa reta. Ao final, utilizamos a média dos 19 valores de $simPorGenero(A, B)$ para designar a "suavidade" da reta.

Para implementar esta métrica, realizamos o seguinte experimento: foram geradas 100 retas aleatórias, nos mergulhos de 2, 3, 5, 7 e 10 dimensões. O resultado pode ser verificado na Figura 2(a), que ilustra um gráfico do tipo "violino", que mostra uma aproximação da distribuição das similaridades por gênero entre pontos consecutivos das 100 retas aleatórias, para cada dimensionalidade do mapa. É possível verificar que o gradiente de gêneros em pares consecutivos de pontos fica cada vez mais suave com o aumento da dimensionalidade do espaço utilizado. Isso serve como indicação de que conjuntos de itens com gêneros similares são agrupados nos mapas multidimensionais, sendo o agrupamento mais coeso em espaços com maior dimensionalidade.

Predominância de um gênero em vizinhanças locais: Selecionamos aleatoriamente 50 itens em cada mergulho e verificamos a distribuição de gêneros em suas vizinhanças locais de tamanhos variados. Para cada tamanho de vizinhança, medida em número de itens contidos na mesma, computamos a similaridade por gênero média entre o ponto central e cada um dos seus vizinhos, usando a definição (2).

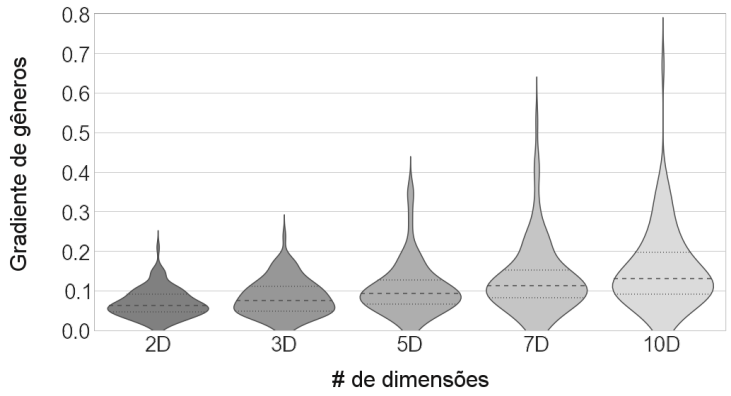
Na Figura 2(b) analisamos a similaridade por gênero entre um item e seus vizinhos no espaço Euclidiano. Mais especificamente, podemos ver a média e intervalo de confiança da distribuição de similaridade por gênero dos itens na vizinhança local, com 95% de nível de confiança. Podemos observar que em vizinhanças de tamanhos crescentes, a similaridade por gênero entre um ponto e seus vizinhos é maior em vizinhanças menores e tende a decair com o aumento da vizinhança. Além disso, observamos que este comportamento da similaridade por gênero ser maior quanto menor o tamanho da vizinhança se torna mais evidente e característico com o aumento da dimensionalidade do espaço.

Análise manual de vizinhanças locais: Escolhemos manualmente alguns filmes e programas de TV de referência, e coletamos os 9 pontos mais próximos deles nos mergulhos de 2, 3, 5, 7 e 10 dimensões. Usamos a opinião subjetiva dos autores, com a ajuda de uma especialista externa que mantém um blog sobre séries de televisão[7], para analisar qualitativamente o que acontecia com essas vizinhanças à medida que a quantidade de dimensões no mergulho era alterada.

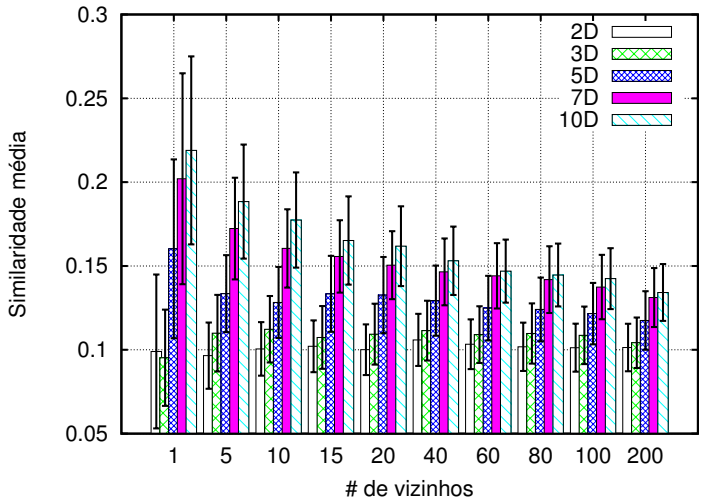
A seguir discutimos os resultados obtidos através da análise manual da vizinhança de alguns conteúdos escolhidos. Foram escolhidos 4 programas de TV e 1 filme: 2 séries de TV, 1 filme de ação e 2 *reality shows*:

(a) Série de Comédia - Parks and Recreation: A Tabela 2 mostra as vizinhanças da série para 2, 5 e 10 dimensões. Para a dimensão 2, alguns filmes de animação

[7] `www.facebook.com/tvlandbr`



(a) Gradiente de gêneros em retas aleatórias.



(b) Semelhança por gênero em vizinhanças locais.

Figura 2. Análise da qualidade dos mergulhos em função da dimensionalidade.

encontram-se na vizinhança, como Ice Age e Shrek (note que ambos pertencem ao mesmo gênero). Entretanto, ao aumentarmos as dimensões, quase toda a vizinhança é formada por séries de comédia, como It's Always Sunny in Philadelphia, 30 Rock, Weeds e Community. Esses títulos são mais representativos do público de Parks and Recreation em comparação aos dos relacionados considerando as outras dimensões, dado que as dimensões menores possuem alguns ruídos como Star Trek, um filme de ficção científica, ou Super 8, um filme de suspense produzido por Spielberg.

(b) Série Teen - Vampire Diaries: Os resultados para a análise da vizinhança desta série é mostrada na Tabela 3. Podemos observar que, para a dimensão 2, aparecem várias séries e filmes mais gerais, como Gey's Anatomy e CSI, enquanto, à medida que as dimensões aumentam, itens mais relacionados com a temática da série Vampire Diaries fazem parte da vizinhança (p.e, Twilight é um filme adolescente sobre Vampiros).

Parks and Recreation		
2D	5D	10D
Ice Age	Cosby Show	Its always sunny in philadelphia
Californication	Conan	Sons of Anarchy
Super 8	Wonder Years	Weeds
Daria	Its Always Sunny in Philadelphia	Community
Knocked up	Golden Girls	Conan
Golden Girls	Boardwalk Empire	Curb your Enthusiasm
Angel	Frasier	Anthony Bourdain no reservations
Its Always Sunny in Philadelphia	I Love Lucy	30 Rock
Shrek Third	Star Trek next Generation	Breaking Bad

Tabela 2: Vizinhança: Parks and Recreation.

Vampire Diaries		
2D	5D	10D
Hells Kitchen	Twilight Saga part 1	Twilight Saga part 1
Sex City	Smallville	Twilight
CSI Crime SceneI nvestigation	Twilight	Greys Anatomy
Greys Anatomy	CSI NY	Xfactor
White Collar	notebook	csi_miami
tangled	CSI Miami	Pretty Little Liars
Notebook	Smurfs	Pirates of Caribbean on Stranger Tides
21 Jump Street	Tangled	CSI NY
Mentalist	Greys Anatomy	Smurfs

Tabela 3: Vizinhança: Vampire Diaries.

(c) Filme de Ação - Duro de Matar: A Tabela 4 mostra a mudança da vizinhança com o aumento das dimensões do espaço Euclidiano. Podemos observar que, em duas e cinco dimensões, sua vizinhança está em uma região com alguns filmes de ficção científica, como Alien, District 9 e Back to the Future. Em 10 dimensões, os vizinhos mais próximos são Quinto Elemento, que tem inclusive o mesmo ator (Bruce Willis) como protagonista, seguido por Exterminador do Futuro, Batman e Indiana Jones, que são filmes de ação/aventura bem mais relacionados entre si.

Duro de Matar		
2D	5D	10D
American Beauty	Terminator	Fifth Element
Tron	Seven	Terminator
Terminator	Terminator 2 (judgment day)	Terminator 2 (judgment day)
Back to the Future 3	Back to the Future 2	Batman
V or Vendetta	Alien	Indiana Jones Raiders of Lost Ark
Back to the Future 2	American Psycho	Indiana Jones Last Crusade
Terminator 2 (judgment day)	Fifth Element	Back to the Future 2
Fifth Element	District 9	Indiana Jones Temple of Doom
Silence of Lambs	Batman	Alien

Tabela 4: Vizinhança: Duro de Matar.

(d) Reality Show de Competição de Moda - Project Runway: Observando a vizinhança deste programa, conforme mostrada na Tabela 5, notamos que para 2

dimensões, a vizinhança tem poucos programas relacionados: os únicos reality shows são Top Chef, que, de fato, é uma competição e compartilha o público de Project Runway, e Pawn Stars, um reality show sobre uma loja de penhores em Las Vegas. No entanto, para 10 dimensões, além do Top Chef ser o ponto mais próximo no mapa, outros reality shows relacionados aparecem, como So you think you can dance, Amazing Race e Americas' next Top Model. Este resultado mostra que o aumento das dimensões incluiu o programa num nicho que, segundo nossa especialista, atende a um público específico e bem definido.

Project Runway		
2D	5D	10D
Hey Arnold	Sweet Home Alabama	Top Chef
Brave	White Collar	Sex City
Crazy Stupid Love	Help	So you think you can dance
Top Chef	Cold case	Americas Next Top Model
Men in Black 3	Cougar Town	Late Night with Jimmy Fallon
Hawaii five0	Top Chef	Raising Hope
Pawn Stars	Private Practice	Beverly Hills 90210
Blind Side	Happy Endings	Amazing Race
Friends with Benefits	So you think you can dance	Masterchef

Tabela 5: Vizinhança: Project Runway.

(e) Reality Show de Competição de Artistas - American Idol: Analisando a vizinhança deste programa, mostrada na Tabela 6, podemos observar que, em 2 dimensões, apesar de o ponto mais próximo ser o The Voice, que é um programa muito parecido com American Idol, os outros programas não tem muito em comum além, talvez, do público alvo. Aumentando o número de dimensões, aparecem cada vez mais reality shows, como Dancing with the Stars, X Factor e America's Got Talent, e a vizinhança passa a fazer mais sentido tanto em público alvo quanto em nicho de mercado.

American Idol		
2D	5D	10D
The voice	The voice	The voice
Snow white Huntsman	Revenge	Dancing with Stars
2 broke girls	CSI NY	Ellen Degeneres Show
Twilight Saga part 1	Xfactor	Twilight Saga part 1
puss_in_boots	Twilight Saga part 1	Pretty Little Liars
Twilight	CSI Miami	Xfactor
Desperate Housewives	New Girl	Greys Anatomy
Wipeout	NCIS los Angeles	Americas got Talent
NCIS los Angeles	CSI Crime Scene Investigation	Vampire Diaries

Tabela 6: Vizinhança: American Idol.

Em suma, através da análise de vizinhanças locais dos itens apresentados é possível observar que as vizinhanças em 2 dimensões se mostraram muito genéricas. No entanto, à medida que o número de dimensões aumenta, as vizinhanças passam a fazer mais sentido, entrando cada vez mais nos nichos específicos de público que cada título atrai. Em muitos dos casos analisados em 2 dimensões, foi possível encontrar

alguma relação superficial entre títulos que poderiam ser interpretados como ruído de sobreposição (ex. gênero ou época de lançamento semelhantes).

5. Trabalhos Relacionados

TV, cinema e redes sociais online: Assistir a programas de TV e a filmes deixou de ser uma atividade restrita as casas e aos cinemas. Atualmente, as pessoas compartilham e discutem sobre estas mídias, individualmente ou em grupo, através de diferentes dispositivos e das redes sociais [Bondad-Brown et al. 2012]. Autores em [Narasimhan and Vasudevan 2012] analisam a viabilidade de usar as atividades em redes sociais (Twitter) para caracterizar o comportamento da audiência. Torrez-Riley [Torrez-Riley 2011] apresenta a evolução histórica do papel da televisão nas interações sociais. Basapur *et. al* [Basapur et al. 2012] descreve o desenvolvimento de uma ferramenta (FanFeeds) para motivar a utilização de outros dispositivos (smartphones, tablets) como extensão da TV e cinema.

Considerando o entendimento do comportamento das pessoas em relação às mídias de televisão e cinema, muitos dos trabalhos encontrados na literatura focam em redes sociais genéricas, principalmente no Twitter [Lochrie and Coulton 2012, Narasimhan and Vasudevan 2012]. Apesar do grande conjunto de trabalhos que estudam o comportamento social e as características estruturais de redes sociais online, não foram encontrados trabalhos que utilizam redes sociais direcionadas aos fãs de TV e cinema, como por exemplo o tvtag.

Medidas de Similaridade: Existem diversas estratégias para obter informação sobre o grau de similaridade entre diferentes mídias. Em relação ao conteúdo de TV e filmes, podemos citar a análise de metadados [Kim et al. 2013] e relações semânticas (baseadas em filtragem de conteúdo) [Yamada et al. 2012, Clinchant et al. 2011, Kim et al. 2013] e filtragem colaborativa [Zhang et al. 2013]. Além destas estratégias, podemos citar estratégias híbridas, que utilizam, por exemplo, filtragem colaborativa e filtragem de conteúdo [Melville and Sindhwani 2010].

Medidas de similaridades baseadas em filtragem colaborativa tipicamente exploram informação disponível ao público e, portanto, são mais escaláveis. Grande parte dos trabalhos exploram o domínio da música [Ragno et al. 2005, Goussevskaia et al. 2008]. Destacamos também [Moore et al. 2012] que apresentam uma metodologia de mergulho e predição de listas musicais baseada em *playlists* de usuários. Considerando o domínio de programas de TV, em [Zhang et al. 2013] os autores se baseiam em microblogs que podem revelar preferências dos usuários e as interações entre os mesmos. A partir da análise realizada, os autores propõem uma estratégia para sistemas de recomendação. Nosso trabalho se difere dos demais por utilizar uma rede social online direcionada aos fãs de TV e cinema para obter as informações que são utilizadas na definição da similaridade entre conteúdos.

Navegação em coleções de mídia: Em relação a navegação de coleções de mídia, muitos trabalhos estão orientados ao domínio da música [Knees et al. 2006, Neumayer et al. 2005, Goussevskaia et al. 2008]. Os artigos relacionados a programas de TV focam, principalmente, na definição da similaridade entre conteúdos. Para navegação, podemos citar a ferramenta proprietária Watchmi [Watchmi 2014] que fornece soluções para navegação e descoberta de conteúdo baseadas em similaridade de conteúdo,

utilizando metadados, preferências e redes sociais (Twitter, Facebook).

6. Discussão dos Resultados e Conclusões

Trabalhar com um mergulho ao invés de um grafo apresenta vantagens de desempenho computacional, tanto em consumo de memória quanto em tempo de processamento. Além disso, os elementos definidos em um espaço Euclidiano possibilitam várias funcionalidades interessantes, como trajetórias, volumes e a noção de direção, que podem ser exploradas na construção de novos aplicativos.

Um bom mergulho coloca os filmes, séries e programas no espaço Euclidiano de forma que títulos similares se encontrem próximos. Portanto, regiões no espaço podem ser associadas a certas propriedades dos itens, tais como gênero, época de lançamento ou público alvo, o que pode ser utilizado para diversos fins. Quando a coleção pessoal de um usuário é mapeada neste "espaço de mídia", as regiões que ela ocupa podem ser representadas como um volume (ou união de vários volumes). De maneira similar, um volume pode ser usado para definir a região de interesse de um usuário, que pode ser explorada para encontrar novos itens que esse usuário ainda não conhece.

Trajetórias, por outro lado, permitem "navegar" suavemente entre programas e regiões. O senso de direção pode ser usado para extrapolar tais trajetórias. Dado um par de filmes, podemos estabelecer uma lista de filmes que apresentem uma transição entre esses títulos. Dada uma lista de filmes, podemos definir maneiras para continuar e estender essa lista, explorando a relação entre os títulos para manter a coerência da lista.

Um mapa pode também abrir caminho para inovações dentro de casa. Ao invés de selecionar um filme navegando em menus textuais e lineares como a lista do netflix, ou percorrendo algumas estantes de DVDs em ordem alfabética em uma loja, os usuários poderão direcionar sua experiência de mídia com instruções de alto nível, como "procure um suspense engraçado", "essa série não, e nem nada parecido com ela", ou "vá em direção ao gênero *comédia francesa*" - e isso poderia retornar uma indicação tão certeira quanto uma conversa com o dono da locadora de filmes da esquina.

Agradecimentos: Ao CNPq, FAPEMIG e Lei de Informática (Projeto LG).

Referências

Basapur, S., Mandalia, H., Chaysinh, S., Lee, Y., Venkitaraman, N., and Metcalf, C. (2012). Fanfeeds: Evaluation of socially generated information feed on second screen as a tv show companion. In *Proceedings of the 10th European Conference on Interactive Tv and Video*, EuroiTV '12, pages 87–96.

Bondad-Brown, B. A., Ricea, R. E., and Pearce, K. E. (2012). Influences on tv viewing and online user-shared video use: Demographics, generations, contextual age, media use, motivations, and audience activity. *Journal of Broadcasting & Electronic Media*, 56:471–493.

Clinchant, S., Ah-Pine, J., and Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 44:1–44:8, New York, NY, USA. ACM.

de Silva, V. and Tenenbaum, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems 15*, pages 705–712.

Goussevskaia, O., Kuhn, M., and Wattenhofer, R. (2008). Exploring music collections on mobile devices. In *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '08, pages 359–362.

Harel, D. and Koren, Y. (2002). Graph drawing by high-dimensional embedding. In *Revised Papers from the 10th International Symposium on Graph Drawing*, pages 207–219.

Kim, J.-M., Shin, P., Kim, J.-J., and Chung, H.-S. (2013). TV Program Retrieval System based on the Computation of Semantic Correspondence. *Journal of Next Generation Information Technology*, 4(8).

Knees, P., Schedl, M., Pohle, T., and Widmer, G. (2006). An Innovative Three-Dimensional User Interface for Exploring Music Collections Enriched with Meta-Information from the Web. In *ACM Multimedia*, pages 17–24.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional scaling*, volume 11. Sage.

Kuhn, M. and Wattenhofer, R. (2007). The theoretic center of computer science. *SIGACT News*, 38(4):54–63.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

Lochrie, M. and Coulton, P. (2012). Sharing the viewing experience through second screens. In *Proceedings of the 10th European Conference on Interactive Tv and Video*, EuroiTV '12, pages 199–202, New York, NY, USA. ACM.

Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., and Ishizuka, M. (2007). Polyphonet: An advanced social network extraction system from the web. *Web Semant.*, 5(4):262–278.

Melville, P. and Sindhwani, V. (2010). Recommender systems. In *Encyclopedia of Machine Learning*, pages 829–838.

Moore, J. L., Chen, S., Joachims, T., and Turnbull, D. (2012). Learning to embed songs and tags for playlist prediction. In *In ISMIR*, pages 349–354.

Narasimhan, N. and Vasudevan, V. (2012). Descrambling the social TV echo chamber. *Proceedings of the 1st ACM MCSS '12*, page 33.

Neumayer, R., Dittenbach, M., and Rauber, A. (2005). PlaySOM and PocketSOMPlayer, Alternative Interfaces to Large Music Collections. In *ISMIR*, pages 618–623.

Ragno, R., Burges, C. J. C., and Herley, C. (2005). Inferring similarity between music objects with application to playlist generation. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, MIR '05, pages 73–80.

Raje, N. (2014). Social tv and the "second screen". `http://trivone.com/blog/social-tv-and-the-second-screen-2/`.

Sareen, H. (2014). Why second-screen media experiences need to be social. `http://www.clickz.com`.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319.

Torrez-Riley, J. (2011). The social tv phenomenon: New technologies look to enhance televison role as an enabler of social interaction.

Watchmi (2014). Funke digital tv guide. `http://corp.watchmi.tv/about/company`.

Yamada, I., Miyazaki, M., Sumiyoshi, H., Matsui, A., Furumiya, H., and Tanaka, H. (2012). Measuring the similarity between tv programs using semantic relations. In *Proceedings of COLING 2012*, pages 2945–2960.

Zhang, Y., Chen, W., and Yin, Z. (2013). Collaborative filtering with social regularization for tv program recommendation. *Know.-Based Syst.*, 54:310–317.