METODOLOGIA PARA RECOMENDAÇÃO DE VÍDEOS BASEADA EM DESCRITORES DE CONTEÚDO VISUAIS E TEXTUAIS

METHODOLOGY FOR VIDEO RECOMMENDATION BASED ON DESCRIPTORS VISUAL CONTENT AND TEXTUAL

Felipe Leandro Andrade da Conceição
Flávio Luis Cardeal Pádua
Adriano César Machado
Anísio Mendes Lacerda
Daniel Hasan Dalip

Resumo: Este trabalho aborda o desenvolvimento de uma nova abordagem multimodal para lidar com o problema de *cold-start* em sistemas de vídeo de recomendação. Dado um documento de vídeo, que geralmente consiste de uma sequência de imagens, áudio e informações relacionadas – como o título, *tags* (rótulos) e descrição –, a recomendação de vídeos visa encontrar uma lista dos vídeos mais relevantes para direcionar os usuários. O problema de *cold-start* é ocasionado quando, ao recomendar um item, não se possui informações sobre o usuário (*cold-start user*) ou sobre o item (*cold-start item*). Mais especificamente, nossa abordagem considera duas modalidades de informação: (i) visual (informações de cor e facial) e (ii) textual (tags, título e descrição) de informação. Para avaliar a recomendação realizada, estendemos o método *Sparse linear method with side information* (SSLIM), que considera as modalidades de informação acima mencionadas e o comparamos com o KNN que aqui representa o *baseline* a ser superado. Neste trabalho, propomos uma abordagem que explora a informação visual e textual para descrever adequadamente vídeos no nosso sistema de recomendação. Os resultados experimentais obtidos em uma coleção real de vídeos, com 207.154 vídeos do Youtube mostra que foi possível melhorar até 7%, quando comparado com uma única modalidade e de 13% sobre o *baseline*, demonstrando a eficácia da abordagem proposta e destacando a utilidade da informação multimodal quando se lida com o problema de *Cold-Start*.

Palavras-chave: Sistemas de Recomendação. Vídeos Online. Web. Cold Start. Abordagem Multimodal

Abstract: This work addresses the development of a new multimodal approach to deal with the cold-start problem in video recommender systems. Given a video document, which usually consists of a sequence of images, audio and related information (such as title, tags and description), video recommendation is formulated as finding a list of the most relevant videos to target users. The cold-start problem is caused when there is no information about the user (cold-start user) or the item (cold-start item) when recomending an item. More specifically, our approach considers two information modalities: (i) visual (color and facial information) and (ii) textual (tags, title, and description) information. To qualify the recommendation performed, we extend a Sparse linear method with side information (SSLIM), which takes into account the aforementioned information modalities. In this work, we propose an approach that exploits visual and textual information to properly describe videos in our recommender system. Experimental results obtained on a real dataset containing 207.154 videos from Youtube shows that we could improve up to 7% when comparing to a single modality model and up to 12% when compared to the baseline, demonstrating the effectiveness of the proposed approach and highlighting the usefulness of multimodal information when dealing with the cold-start problem.

Keywords: Recommender Systems. Online Video Systems. Web. Cold Start. Multimodal Approach

1 - INTRODUÇÃO

A *World Wide Web* (WWW ou Web) tornou-se, ao longo dos anos, parte integrante da vida quotidiana de milhões de pessoas, na qual os usuários podem, por exemplo, comprar produtos e serviços, visualizar e compartilhar fotos, assistir vídeos e ler jornais (MA et al., 2005). Entretanto, devido a grande quantidade de objetos informacionais, estes usuários encontram dificuldade em filtrar conteúdo útil para sua necessidade de informação. Nesse cenário, sistemas de informação têm incorporado métodos de recomendação para personalizar a disponibilização de conteúdo para esses usuários (RESNICK e VARIAN, 1997).

Sistemas de recomendação surgiram com o objetivo de auxiliar pessoas (usuários) a encontrar itens que sejam do seu interesse dentre uma grande e crescente variedade de opções. Tais sistemas modelam o perfil dos usuários e seu histórico de utilização do sistema com o objetivo de encontrar itens relevantes. Para isso, são aplicadas técnicas como

mineração de dados e aprendizagem de máquina, que visam mapear os interesses de cada usuário do sistema.

Sistemas de recomendação podem ser encontrados em diversos domínios. Muitos sítios conhecidos utilizam estes sistemas para recomendar itens para seus usuários. O *Youtube*, que é um dos maiores sistemas de compartilhamento de vídeos do mundo, utiliza sistemas de recomendação para identificar padrões no perfil dos usuários e por meio dele encontrar novos vídeos que podem ser do seu interesse (DAVIDSON et al., 2010). Outro exemplo de sítio Web que utiliza um sistema de recomendação é a Amazon.com, que utiliza tal sistema para recomendar produtos (ex: livros e músicas) para seus clientes (LINDEN et al., 2003).

Sistemas de recomendação podem ser classificados em três categorias: (1) sistemas baseados em filtragem colaborativa que recomendam itens para o usuário por meio de outros usuários que possuem perfis semelhantes (i.e., que se interessam pelos mesmos itens), (2) sistemas baseados em conteúdo que recomendam itens que possuem conteúdo semelhantes aqueles já consumidos pelo usuário e (3) sistemas híbridos que combinam os dois tipos de recomendação citados acima (BOBADILLA et al., 2013).

Os sistemas baseados em filtragem colaborativa representam os estado-da-arte em sistemas de recomendação. Porém, estes sistemas assumem que exista informação suficiente sobre os usuários e/ou itens para que a recomendação seja relevante. O cenário no qual a não existe informação de preferência por parte dos usuários (ex: novos usuários) ou informação de consumo dos itens (ex: novos itens) é chamado de *cold-start* (BERNARDI et al., 2015). Em outras palavras, *cold-start* impede que técnicas colaborativas sejam utilizadas pois não existe informação de preferências disponível.

Neste trabalho estamos interessados no cenário em que temos informação insuficiente sobre os itens, i.e., *item cold-start*. Para contornar a limitação das abordagens colaborativas, propomos explorar a informação referente ao conteúdo dos itens. Especificamente, exploramos informação visual (ex: cor) e textual (ex: título). Existem várias pesquisas propondo métodos para lidar com *cold-start* em sistemas de recomendação (PARK e CHU, 2009; YAN et al., 2015). Porém, tais trabalhos baseiam-se essencialmente em descritores textuais, tais como, *tags (rótulos)*, títulos e resumos descritivos acerca do conteúdo do objeto informacional.

Neste sentido, este trabalho propõe uma abordagem multimodal, isto é, baseada na combinação de descritores de diferentes modalidades (visual e textual) para a recomendação de vídeos em cenários onde a maioria dos objetos informacionais de um usuário foram consumidos apenas por ele. Nossa hipótese principal é que, utilizando a combinação de diferentes modalidades de descritores, conseguimos melhorar a descrição dos itens e, por consequência, a relevância dos vídeos recomendados.

As principais contribuições deste trabalho são: (1) a apresentação de abordagem multimodal para recomendação de vídeos – tal abordagem conseguiu uma melhoria de 7% em comparação quando utilizamos apenas uma modalidade (sem combiná-la) e de 13% sobre o *baseline*; (2) a proposta de um novo descritor para descrever vídeos baseado na identificação de faces nos vídeos; (3) o estudo do impacto dos diferentes descritores para a tarefa de recomendação no cenário *cold-start*.

O restante deste artigo está organizado da seguinte forma. A Seção 2 descreve os principais trabalhos relacionados sobre a temática abordada neste artigo. Na Seção 3 apresentamos a nossa abordagem de recomendação multimodal, logo após, na Seção 4 apresentamos nossos resultados experimentais que validam a metodologia proposta. Finalmente, na Seção 5 descrevemos as conclusões e algumas direções de trabalhos futuros.

2 – TRABALHOS RELACIONADOS

Na última década, sistemas de recomendação tem motivado diversos trabalhos. Nesta seção descrevemos trabalhos relacionados ao tema deste artigo. Inicialmente, apresentamos os trabalhos cujo objeto de recomendação são vídeos. Em seguida, como esse trabalho combina informações do usuário e do conteúdo do item, apresentamos métodos que fazem tais combinações. Por fim, detalhamos estratégias de recomendação que tratam do problema de *cold-start*.

Sistemas de filtragem colaborativa utilizam o histórico dos usuários e suas indicações no passado para fazer indicação a outros usuários com preferências similares. No contexto de recomendação de vídeos, Davidson et al. (2010) apresentam um método de recomendação de vídeos no Youtube. Para isso eles utilizam quais vídeos o usuário consumiu e/ou indicou como tendo gostado. Então, eles utilizaram técnicas de associação para identificar quais vídeos devem ser recomendados. Em Zhou et al. (2010), os autores demonstraram que a recomendação de vídeos no Youtube é uma das mais importantes

fontes de visitas aos seus vídeos.

Vários aspectos do conteúdo podem ser importantes para melhorar a relevância da recomendação. Por isso, alguns autores têm utilizado uma abordagem híbrida, combinando filtragem colaborativa com baseada em conteúdo. Kim et al. (2010) mostraram uma abordagem de filtragem colaborativa que melhorou a qualidade da recomendação. Nesta abordagem, utilizou-se também informações de tags com o objetivo de auxiliar na descoberta das preferências do usuário e, além disso, auxiliar quando se tem pouca ou nenhuma informação sobre o usuário (i.e., cold-start user). No presente trabalho estudaremos também o uso de *tags* para auxiliar na recomendação de vídeos.

Além disso, Bobadilla et al. (2012) mostram que pode-se utilizar a importância de um determinado item para ponderar e melhorar a recomendação. Esse trabalho mostrou que é possível utilizar a importância através do item mais vendido, visualizado e/ou comentado, por exemplo. Outro aspecto foi analisado por Deng et al. (2015), onde utilizaram os tópicos (populares no momento) de interesse para o usuário para recomendar vídeos do Youtube que foram compartilhados pelo Twitter. Yan et al. (2015) também usaram informações do usuário no Twitter para enriquecer a recomendação e auxiliar principalmente quando não se sabe muito sobre a preferência do usuário (i.e., *user cold-start*). Park et al. (2009) tratam o problema de *item cold-start*. Os autores utilizaram as seguintes características para recomendação de filmes: o gênero, o elenco, entre outros aspectos do conteúdo para inferir a popularidade e, assim, utilizaram tais informações quando não se sabe nada sobre este item novo. De forma similar, utilizando informações demográficas e histórico do usuário, também tratam o problema de *user cold-start*.

Li et al. (2014) realizaram recomendação multimodal de vídeos baseada em características de áudio e imagem (usando as cores). Eles utilizaram o classificador SVM (VAPNIK, 1995) e modelaram o problema como um problema de classificação binária, assim, um item pode ser classificado como relevante ou não-relevante para o usuário. Além disso, eles utilizaram a distância do cosseno como métrica para identificar quais vídeos podem ser considerados similares. Ele também trata o problema do novo item, inferindo se ele será recomendado ou não através do modelo criado pelo classificador SVM.

Nosso trabalho é similar a este, porém, diferentemente de Li et al. (2014), aqui adotamos uma abordagem não supervisionada, ou seja, a anotação manual prévia de quais itens são relevantes não é necessária. Em Li et al. (2014), uma vez que é utilizado o SVM, a

anotação é condição necessária. Além disso, nesse trabalho utilizamos, além da cor, a similaridade de faces presentes nos vídeos. Através de nossos experimentos, conseguimos demonstrar que a utilização do descritor de faces é importante e consegue aprimorar a precisão para a recomendação de vídeos.

No presente trabalho, também iremos combinar dados de preferência do usuário com outras informações. Para isso, usaremos o SSlim (Zheng et al., 2014), que combina duas matrizes para realizar as recomendações. O método sugere combinar informação colaborativa, caso exista, com informação de conteúdo dos vídeos para realizar a recomendação. Assim, diferentemente das abordagens citadas acima, exploramos características específicas do domínio de recomendação de vídeos (ex: cores, faces, etc) para recomendá-los aos usuários.

3 - ABORDAGEM PROPOSTA

Nesta seção, descrevemos nossa abordagem multimodal para recomendação, i.e. baseando-se em conteúdo visual e textual dos vídeos. Quanto à representação textual, uma vez que vídeos são acompanhados pelo seu título, *tags* e descrição utilizaremos tal informação para representá-los. Quanto ao conteúdo visual, utilizaremos a representação através do histograma de cores e faces.

Dado que utilizamos o conteúdo dos vídeos para recomendá-los, precisamos de utilizar métricas para calcular a similaridade entre eles. Assim, primeiramente, na Seção 3.1 apresentamos como foi feita a similaridade textual e, na Seção 3.2, é apresentado como foi realizada a similaridade visual. Finalmente, na Seção 3.3 apresentamos a abordagem proposta.

É importante ressaltar que, apesar deste trabalho apresentar um novo método de recomendação de vídeos, tal método pode ser aplicado em outros domínios. Por exemplo, em recomendação de produtos podem ser utilizados textos que descrevem os produtos tão bem como uma imagem descritiva do mesmo.

3.1 – Similaridade Textual

Para calcular a similaridade textual, utilizamos a representação *Bag of Words* (BOW). Assim, cada vídeo é representado pelo conjunto de palavras que o descreve. Cada palavra pode possuir uma importância diferente para descrever o vídeo. Assim, nesta representação,

definimos que cada vídeo como um vetor de pesos $w_{pi}(w_{i1}, w_{i2}, \dots, w_{in})$ onde w_{ip} é o peso da palavra *p* no vídeo *i*. Após calcular o peso de cada palavra, é possível calcular a similaridade entre dois vídeos v_i e v_j por meio da distância do cosseno.

Para calcular o peso w_{ip} utilizamos métrica TF-IDF (BAEZA-YATES e RIBEIRO-NETO, 2011) de cada palavra, como na equação abaixo:

$$w_{pi} = TF_{pi} \times IDF_p$$

onde TF_{pi} é a frequência do termo (do inglês, *term frequency*) calculada usando a frequência f_{pi} da palavra *p* no vídeo *i* através da seguinte equação: $1 + \log_2(f_{pi})$. O IDF_p representa o inverso da frequência da palavra *p* calculado a partir da equação $\log_2(\frac{N}{n_p})$. Onde *N* é a quantidade total de vídeos na coleção e n_p é a quantidade de vídeos em que a palavra *p* aparece. Esta métrica é relacionada com quão discriminativa é uma palavra, ou seja, quanto mais vídeos a palavra *p* aparecer, menor será seu *IDF*.

Assim, a partir de agora cada vídeo é representado por um vetor de pesos. Dessa forma, para calcularmos a similaridade textual entre dois vídeos v_i e v_j ($simTxt(v_i, v_j)$ utiliza-se a distância de cosseno, conforme Equação 1, ou seja, $simTxt = cosseno(v_i, v_j)$.

$$cosseno(v_i, v_j) = \frac{v_i . v_j}{||v_i|| \, ||v_j||} \quad (1)$$

A similaridade dos vídeos varia de 0 a 1 e quanto mais próximo de 1, mais similar são os vídeos considerados.

3.2 - Similaridade Visual

Para fazermos a similaridade visual, o vídeo é representado como $v_i = (q_1, q_2, \dots, q_n)$ onde q_k é o *k*-ézimo quadro do vídeo *i*, composto por *n* quadros. Cada quadro *q* é representado pela média dos descritores do histograma de cores vermelho, verde, azul, matiz, saturação e brilho (i.e. RGB e HSV). Outra representação de um vídeo abordado neste trabalho é a utilização das faces dos participantes de cada vídeo, desta forma para cada quadro q_k do vídeo são identificadas todas as faces para compor uma segunda representação dos vídeos v_i(VIOLA e JONES, 2004).

3.2.1 - Similaridade de Faces

Características faciais são um conjunto de informações que caracterizam a face humana. É possível definir inúmeras destas características, podendo ser representadas, por ex., pela largura da boca, espaço entre os olhos ou tamanho do nariz. Nesse trabalho uma face é representada pelos seus descritores de textura (CHUI, 1992), cor e forma (MING-KUEI, 1962). A intuição quanto ao uso de faces é que um usuário que tem preferência por uma determinada pessoa (ex: um ator ou cantor) pode estar interessado em outros vídeos nos quais esta pessoa apareça.

Similarmente ao conteúdo textual, utilizaremos a representação da *Bag of Faces* (BOF) para representar as faces que estão presentes nos vídeos. De forma análoga à representação textual, cada vídeo é representado por um vetor de pesos $v_i = (f_{i1}, f_{i2}, \dots, f_{in})$ onde f_{ip} é o peso da face *p* no vídeo *i*. A similaridade dos vídeos é feita utilizando a distância do cosseno (Equação 1), ou seja, $simFace(v_i, v_j) = cosseno(v_i, v_j)$.

3.2.1 - Similaridade de Cor

Similarmente ao conteúdo textual, criamos uma matriz de pesos para possibilitar o cálculo da similaridade entre dois vídeos v_i e v_j onde cada peso é o valor médio do descritor de cor de cada quadro do vídeo. Note que a medida que o vídeo possui mais quadros, ele terá mais pesos, porém para usar o PCA, temos que ter vídeos com o mesmo número de pesos.

Assim, com o objetivo de deixar tanto v_i e v_j pelo mesmo número de quadros, seja *m* o número de quadros do vídeo menor, utilizamos apenas os *m* quadros iniciais dos vídeos v_i e v_j para a similaridade.

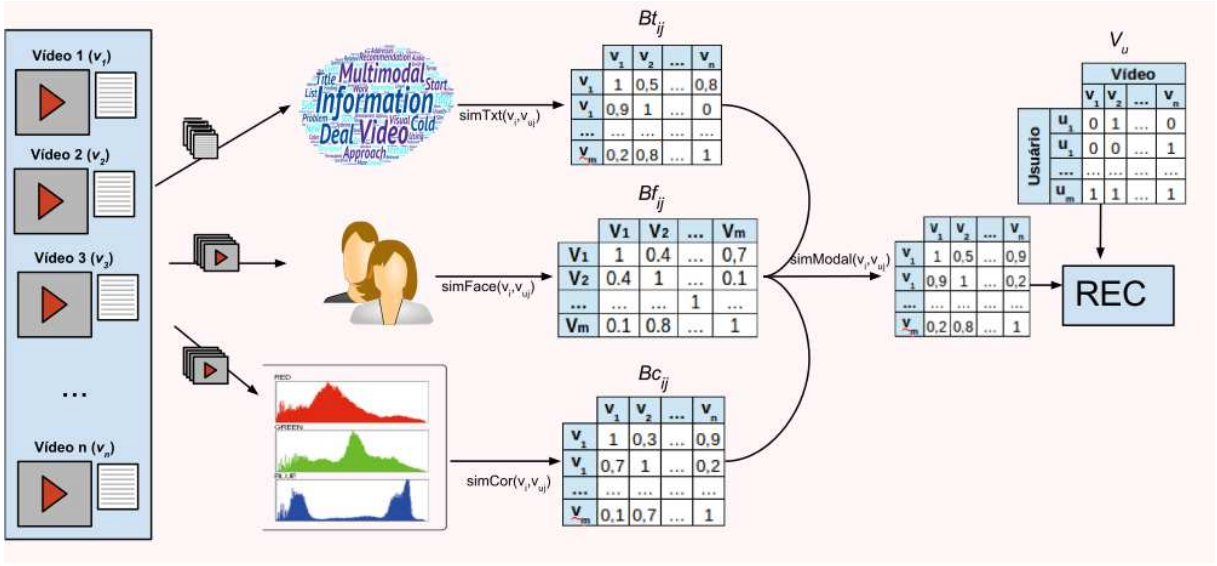
Dessa forma, como cada quadro *q* é representado por 6 cores, podemos obter uma matriz onde w_{ci} é a média do descritor da cor *c* do quadro *q* do vídeo *i*. Os quadros foram colocados de maneira sequencial, ou seja, o quadro *q* é representado por seis pesos w_{ci} onde *c=q* até *c=q+5*. Com o intuito de reduzir o custo computacional associado a esta matriz, foi utilizado a Análise de Componentes Principais (do inglês, PCA – *Principal Component Analysis*) (PEARSON, 1901) para reduzir a dimensionalidade da matriz de descritores visuais dos vídeos.

Assim, considerando que desejamos comparar v_i e v_j, representamos o vídeo como

um vetor $(w_{1i}, w_{2i}, \ldots, w_{mi})$ onde *m* é a quantidade de quadros do menor vídeo entre v_i e v_j. Com tal representação, podemos calcular a similaridade visual entre os vídeos v_i e v_j também por meio da distância de cosseno (cf. Equação 1), ou seja, $simCor(v_i, v_j) = cosseno(v_i, v_j)$.

3.3 - Similaridade Multimodal

Figura 1: Método de Recomendação Multimodal



A visão geral de nosso método é demostrada na Figura 1. Nesta figura é ilustrado o processo de recomendação de vídeos para um determinado usuário *u* utilizando informações visuais e textuais a partir de um conjunto $V = \{v_1, v_2, v_3, \ldots, v_n\}$ de *n* vídeos existentes e o conjunto V_u no qual $V_u \subset V$ de vídeos que o usuário *u* demonstrou interesse. Para cada fonte de informação (i.e., textual e visual), precisamos calcular a similaridade entre elas. A similaridade será uma das entradas do método de recomendação SSLIM (ZHENG et al., 2014) que, por fim, retornará uma lista de vídeos recomendados para cada usuário *u*.

Mais especificamente, a informação visual e textual do vídeo são preprocessadas separadamente para obtermos as suas respectivas representações em *Bag of Words* e histogramas de cores. A partir disso, através da função *simText* (cf. Seção 3.1), é calculada a similaridade textual entre todos os vídeos da coleção, gerando assim a matriz de similaridade Bt_{ij} de tal forma que Bt_{ij} representa a similaridade textual entre os vídeos v_i e v_j. Do mesmo modo, as funções *simCor* e *simFace* calculam a similaridade de cor e facial,

respectivamente (cf. Seção 3.2.2 e Seção 3.2.1). Tais similaridades irão produzir as matrizes Bc_{ij}e Bf_{ij} representando as respectivas similaridades de cor e facial entre os vídeos v_i e v_j.

Logo após, precisamos combinar a similaridade de texto (matriz Bt_{ij}) e facial (matriz Bf_{ij}), além da similaridade de cor (matriz Bc_{ij}). Para isso, usamos o método *simMulti* gerando então a matriz Bm_{ij}, a qual representa a similaridade entre os vídeos v_i e v_j utilizando tanto a representação visual e textual. Para isso, utilizamos o valor máximo entre as similaridades para cada vídeo v_i e v_j, conforme apresentado na Equação 2.

$$simMulti = \max(Bt_{ij}, Bc_{ij}, Bf_{ij}) \quad (2)$$

Por fim, o método SSLIM é utilizado para efetuar a recomendação de vídeos. Para isso, este método utiliza como entrada a matriz de similaridade entre os vídeos Bm_{ij} e, além disso, a matriz de usuário x vídeo A_{uv} onde $A_{uv} = 1$ quando o usuário *u* considera o vídeo *v* relevante e $A_{uv} = 1$, caso contrário. As constantes α e β foram definidas com os valores 0,001 e 5, respectivamente. Esses valores foram definidos a partir de procedimento de calibração do algoritmo. Além disso, note que se quisermos usar apenas as informações textuais ou apenas as informações visuais para recomendação, podemos utilizar as suas respectivas matrizes ($Bt_{ij}, Bc_{ij}\ ou\ Bf_{ij}$), ao invés da matriz Bm_{ij}.

4 - EXPERIMENTOS

Nesta seção discutimos os resultados experimentais, que validam o modelo de recomendação proposta, além de comparar com um *baseline* tradicional. Na Seção 4.1, apresentamos a base de dados utilizada neste trabalho. Na Seção 4.2, apresentamos a metodologia experimental. Por fim, na Seção 4.3 discutimos os resultados obtidos.

4.1 - Coleção de Vídeos

Construímos uma base de dados tendo como referência o *dataset* (CHENG et al., 2008) que possui uma coleção de mais de dois milhões de usuários. Nossa coleção foi feita pela coleta dos metadados dos vídeos presentes em cada lista de execução (do inglês, *playlists*) dos usuários e completamos com metadados, tais como *tag* e descrição, que a princípio não estão presentes em (CHENG et al., 2008). Durante a coleta, foram utilizados usuários que possuem um mínimo de 10 vídeos por lista de execução, desta forma, foram coletados metadados referentes a 10.208 usuários.

Como os vídeos da lista de execução do usuário são de seu interesse, foi considerado que um determinado vídeo *v* é relevante para o usuário *u* caso este vídeo esteja na lista de execução de *u*. Além disso, foram utilizados apenas vídeos com duração inferior a 20 minutos para diminuir a complexidade de tempo de execução dos experimentos.

Para cada vídeo, foi extraído também seu título, descrição e *tags* e realizado preprocessamento textual, no qual as palavras comuns e com pouca relevância semântica (ex: artigos, preposições, etc), i.e. *stopword*, foram removidas. Para remoção das *stopwords* foi utilizada a biblioteca *stopwords* presente no repositório do *Python*. No total, foram extraídos 207.154 vídeos.

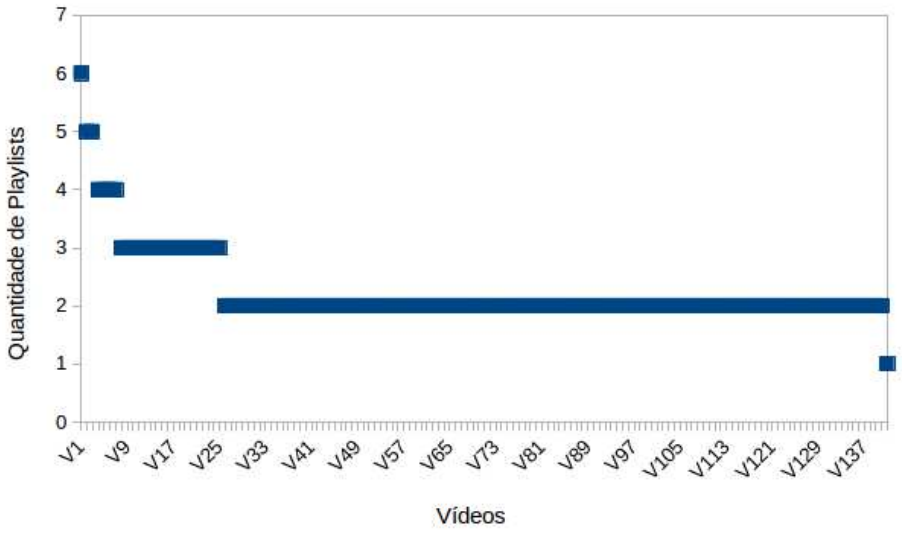


Figura 2: Número de ocorrência de vídeos nas playlists

A Figura *2* apresenta os 200 vídeos que foram adicionados em mais listas de execuções. Como temos mais de 200 mil vídeos, a maioria (99%) dos vídeos dos vídeos dessa base de dados estão em apenas 1 lista de execução e o vídeo com mais ocorrências está em 6 listas de execuções. Por isso, é importante utilizar uma abordagem que leva em consideração o conteúdo e a similaridade entre os vídeos, uma vez que não existe informação colaborativa suficiente.

4.2 - Metodologia Experimental

Nossos experimentos têm como objetivos: (1) analisar o impacto do uso de uma abordagem multimodal na recomendação de vídeos; (2) comparar a abordagem proposta com algum método de referência ou *baseline*; (3) verificar qual é a melhor representação textual (apenas título, *tags* ou descrição) de um item.

O resultado de nosso método é um *ranking* de similaridade para cada usuário *u*. Assim, utilizamos a métrica que calcula de precisão média desse ranking nas primeiras *k*-ésimas posições (*P@k*) da lista de recomendação. Ou seja, dentre os *k* vídeos considerados mais relevantes pelo método, qual é a proporção média de vídeos que são realmente relevantes para determinado usuário *u*. Esta métrica é calculada por meio da seguinte fórmula:

$$P@k = \frac{1}{n} \sum^{u \in U} \frac{hits(u)}{k}$$

onde *U* é o conjunto de todos os usuários, *hits(u)* é o número de vídeos relevantes nas *k*-ésimas primeiras posições do *ranking* de recomendação do usuário *u*.

Nossa abordagem foi avaliada de duas formas: (1) comparando-a com a abordagem que usa apenas uma modalidade (facial, cor ou textual) e (2) comparando-se também com um *baseline* usual na tarefa de recomendação - o método KNN (do inglês, *K-Nearest Neighbours*) (SILVERMAN, 1986).

O KNN é um método de aprendizado de máquina. Neste método, assume-se que se tem acesso a *n* itens previamente rotulados (chamados dados de treino) da seguinte forma $S = \{(v_1, r_1), (v_2, r_2), \dots, (v_n, r_n)\}$ onde, em nosso caso, v_i é um vídeo representado pelo resultado da similaridade de cor, face e texto e r_j é o resultado alvo que, em nosso caso, pode assumir o valor de "sim" ou "não" relevante. Com isso, o KNN irá aprender a combinar a similaridade de cor, facial e textual de um item para assim recomendar itens ao usuário.

Dessa forma, o KNN cria um modelo utilizando os dados em *S* de tal forma que caso haja uma lista de *m* vídeos que precisamos de saber as suas relevâncias $I = \{v_1, v_2, \dots, v_n\}$ para cada item $v_i \in I$ ele retorna sua classe predita (verificando os vizinhos previamente rotulados mais próximos de v_i) e a probabilidade de pertencer a esta classe. Assim, para recomendarmos tais vídeos, utilizamos os itens preditos como relevantes ($r_i = "sim"$) e ordenamos de acordo com a probabilidade de ser relevante.

Os experimentos foram realizados utilizando a metodologia de validação cruzada de

cinco partições (*5-fold cross validation*) (MITCHELL, 1997). Assim, nossa base de dados foi dividida em cinco partes e cada experimento repetido 5 vezes. Em cada repetição, uma partição diferente representando 1/5 da coleção foi avaliada enquanto o restante, no caso do KNN, foi usado para o treino.

4.3 – Resultados

Quadro 1: Combinações dos descritores de texto

Descritor	P@1	P@2	P@3	P@4	P@5
Título	0,679	0,613	0,535	0,467	0,404
Título+Tag	0,604	0,528	0,478	0,396	0,328
Título+Desc	0,650	0,594	0,535	0,462	0,400
Tag+Desc	0,642	0,604	0,547	0,486	0,411
Titulo+Tag+Desc	0,660	0,613	0,560	0,486	0,419

Esta Seção discute os resultados obtidos, avaliando a eficiência da abordagem proposta. Aqui serão abordados os resultados obtidos pela combinação dos descritores de texto e visual. Primeiramente, buscamos descobrir qual é a melhor representação textual que podemos utilizar. Em seguida, comparamos a abordagem multimodal com a utilização de apenas uma fonte de informação. Finalmente, encerramos os resultados comparando nosso trabalho com um *baseline*, que é o método KNN.

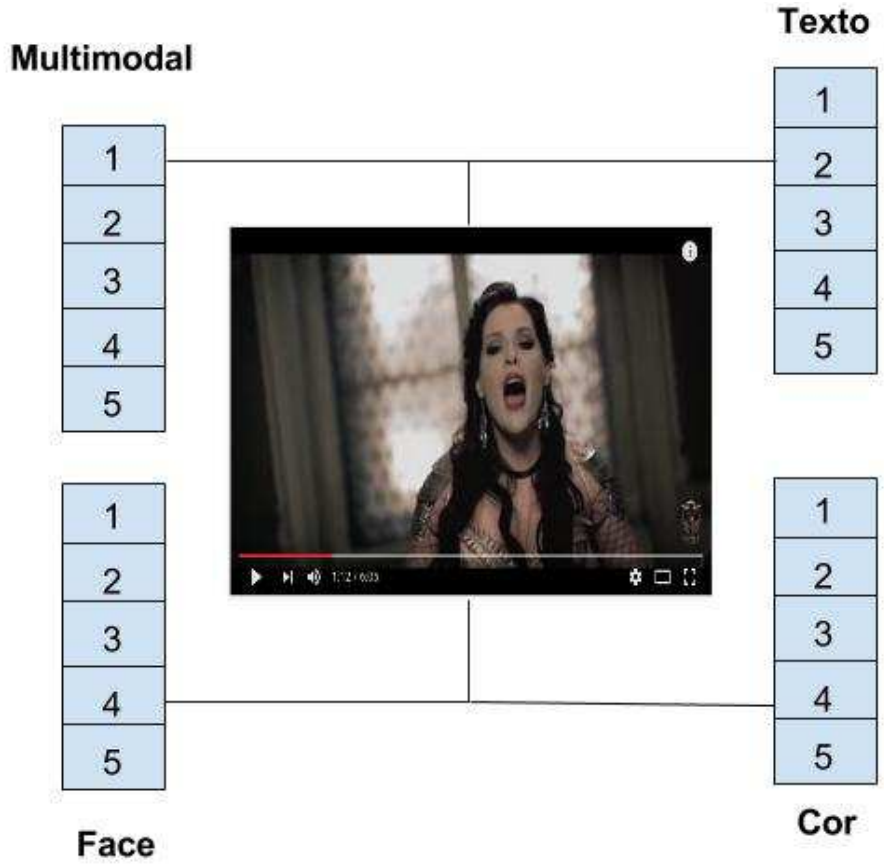
O Quadro 1 apresenta os resultados obtidos utilizando apenas os descritores de texto para a tarefa de recomendação de conteúdo. Para isso, utilizamos todas as combinações possíveis de representações textuais utilizando a descrição, *tag* e o título e utilizamos apenas elas para recomendar conteúdo. Ou seja, neste caso utilizamos apenas a matriz de similaridade textual.

Como se pode perceber, apesar do bom desempenho ao utilizar-se o título descrever o vídeo mais relevante (*P@1*), a combinação de título, *tag* e descrição apresentou um desempenho superior em relação aos outros descritores de texto nas demais posições do *ranking*. Por este motivo, selecionamos o descritor contendo título, *tag* e descrição para os nossos experimentos com a abordagem mutimodal.

Com o objetivo de verificar se é possível melhorar a performance do método através

da combinação multimodal, utilizamos a melhor representação textual e utilizamos nossa abordagem multimodal para combiná-las com a informação visual.

Figura 3: Vídeo

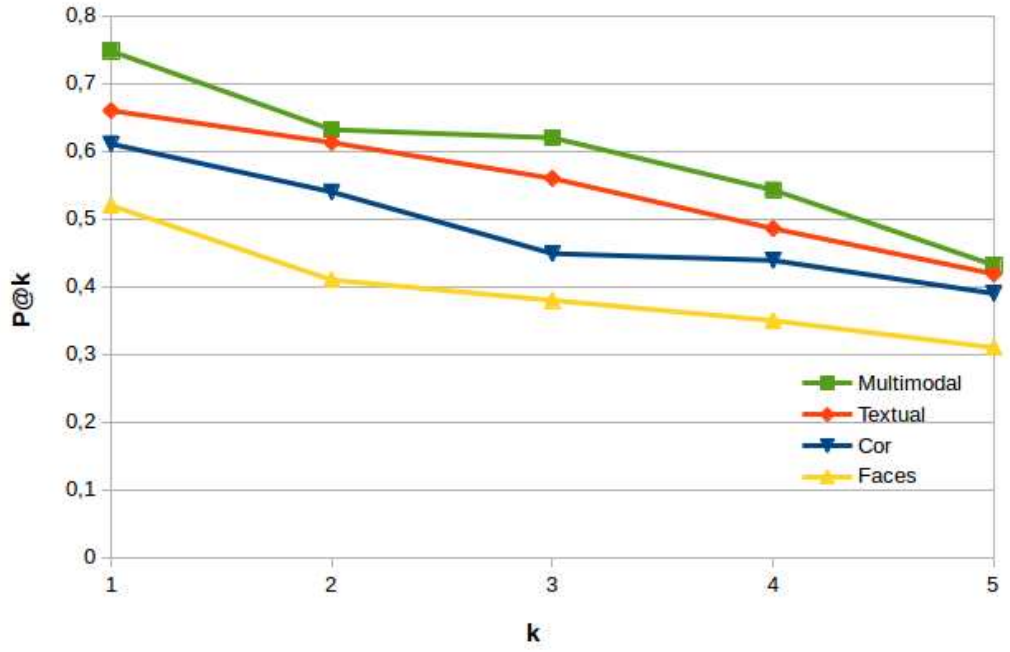


Chama-se a atenção para o vídeo apresentado na Figura 3, onde é feita uma avaliação para recomendação deste item para um usuário. É um item real extraído de nossa base de dados para explicação. Este vídeo possui duração de 6 minutos e recebeu 6.853 curtidas dos usuários que o avaliaram. Os descritores de cor e faces apresentaram uma precisão inferior aos descritores textuais e multimodal, muito provavelmente por não possuir características similares a outros itens da *playlist* deste usuário. O descritor de texto apresentou um bom resultado no ranking porque o conjunto de palavras que descrevem os vídeos presentes na *playlist* deste usuário permitiu uma similaridade melhor entre seus vídeos. Uma vez combinados, os descritores permitiram uma precisão superior e melhor qualificação o vídeo relevante para esse usuário da base de dados, pois muitas lacunas de similaridade deixadas individualmente por cada um foram preenchidas durante a combinação de suas diversas

características na abordagem multimodal proposta.

A Figura 4 apresenta o resultado da métrica de precisão ao utilizar nossa melhor abordagem para as representações textual e visual (cor e faces) isoladas e comparando-as com o resultado de nossa abordagem multimodal nas *k*-ésimas posições do *ranking* (com *k* variando de 1 até 5).

Figura 4: Comparação entre os modelos de recomendação textual, visual e multimodal

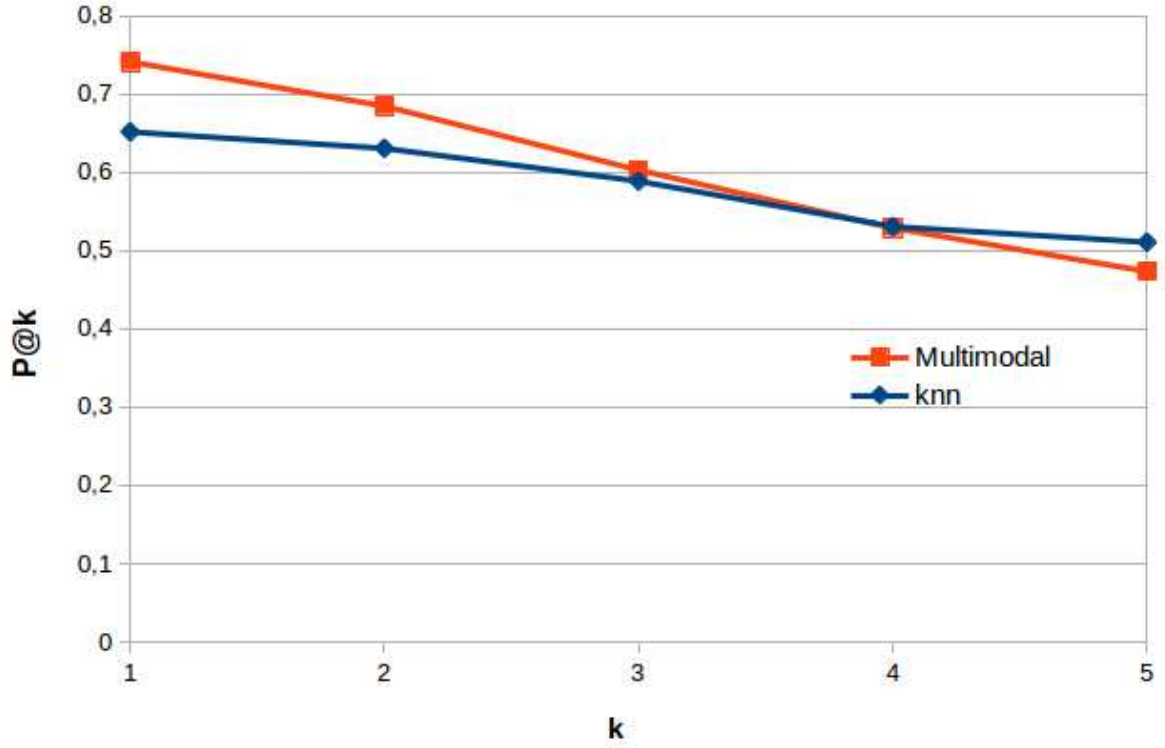


É possível observar que, individualmente o texto consegue a melhor performance. E, quando utilizamos nossa abordagem multimodal, para todas as posições testadas, nossa abordagem conseguiu ser melhor do que utilizando apenas uma fonte de informação. O ganho de precisão da nossa abordagem é de 13% quando *k=1* e 3% quando *k=5*. Assim, podemos observar que utilizando nossa abordagem multimodal é possível melhorar a performance para a tarefa de recomendação, com valores representativos à abordagem unimodal ou com descritor individual.

Na Figura *5* comparamos nossa abordagem com o KNN também utilizando a precisão nas top 1 até top 5 posições. Como podemos perceber, nossa abordagem apresenta uma maior precisão nas 3 primeiras posições, conseguindo uma melhoria de até 8%. Além disso,

cabe ressaltar uma importante vantagem de nossa abordagem proposta em relação ao *baseline* KNN: a nossa abordagem não requer um treino previamente rotulado, o que é uma importante característica para vários cenários de recomendação de itens, como no caso de objetos multimídia.

Figura 5: Comparação entre os modelos multimodal e KNN



Na próxima seção apresentamos uma síntese do trabalho realizado, com as principais conclusões obtidas, as contribuições alcançadas e propostas de trabalhos futuros que pretendemos desenvolver.

5 - CONCLUSÕES

Existe uma variedade de métodos propostos para recomendação de itens para diferentes cenários, incluindo conteúdo multimídia (p. ex., vídeos). Porém existem muitos desafios para garantir a eficácia dos métodos de recomendação. Em especial um dos problemas clássicos diz respeito ao problema conhecido como *item cold-start*, que refere-se à dificuldade de recomendação de itens novos que surgem no sistema e para os quais não existe avaliação de qualidade e relevância.

Este trabalho propôs e validou uma abordagem multimodal para recomendação que

explora o conteúdo de vídeos para minimizar os efeitos do problema de *item cold-start*. Especificamente, apresentamos um método baseado na combinação de descritores de diferentes modalidades (visual e textual) para a recomendação de vídeos em cenários onde a maioria dos objetos informacionais de um usuário foram consumidos apenas por ele, ou seja, onde existe o desafio do *item cold-start* já explicado. Portanto, a premissa em que se baseia nossa proposta é de que a combinação de diferentes modalidades de descritores possibilita melhorar a qualidade da tarefa de recomendação considerando o problema de *item cold-start*.

Para fins de validação e análise desta abordagem, foram utilizados dados reais, que consistem de vídeos do Youtube, que é um dos maiores sistemas do mundo de compartilhamento e distribuição de vídeos. Assim, conseguimos demonstrar que é possível combinar conteúdo textual, cor e faces para assim conseguir uma melhoria na precisão quando comparado à utilização de uma única fonte de informação. Além disso, comparamos nosso método a um *baseline* que também realiza esta combinação, mas utilizando um método de aprendizado de máquina.

A partir dos experimentos, nos quais observamos melhoria na precisão da recomendação, concluímos que a utilização de faces contribuiu para a recomendação de vídeos quando comparada com aos modelos unimodais e também com outras propostas multimodais sem utilização de faces.

As principais contribuições alcançadas com este trabalho foram as seguintes:

- A apresentação de abordagem multimodal para recomendação de vídeos, que alcançou ganhos representativos, com melhora de até 7% da precisão média em comparação aos modelos unimodais e de até 13% sobre o *baseline*;
- A proposta de um novo descritor para auxiliar na recomendação de conteúdo por meio de identificação facial, que poderá ser aprimorado e aplicado a outros cenários;
- A identificação de quais descritores textuais são melhores para esta tarefa. Foi identificado que quando se usa todas as informações, como a descrição do vídeo, o título e *tags*), é possível obter melhor qualidade nas recomendações.

Como trabalhos futuros, pretende-se aplicar os modelos em outras bases de dados de conteúdo multimídia, por exemplo de programas televisivos que são disponibilizados na Web. Nosso modelo pode prover melhor qualidade na recomendação aos usuários, bem como ser utilizado para recomendações complementares, como em propaganda

disponibilizada em vídeos na Web.

Além disso, pretendemos gerar novas versões do modelo multimodal de recomendação, agregando outras características visuais dos objetos multimídia e investigando quais atributos podem agregar valor, produzindo melhor taxa de precisão e também possibilitando prover maior diversidade e novidade nas recomendações, que são outras métricas que podem ser relevantes, dependendo da aplicação.

REFERÊNCIAS

RESNICK, P.; VARIAN, H. R. **Recommender systems**. Communications of the ACM , v. 40, n.3, p. 56-58, 1997.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. **Modern Information Retrieval - the concepts and technology behind search**, Second edition . Pearson Education Ltd., Harlow, England, 2011.

BERNARDI L.; KAMPS, J.; KISELEVA, J.; MÜLLER, M. J. I.. **The continuous cold start problem in e-commerce recommender systems**. CoRR , 2015.

BOBADILLA, J.; HERNANDO, A.; ORTEGA, F.; GUTIERREZ, A.. **Collaborative Filtering based on significances**. Information Sciences , v. 185, n. 1, p. 1-17, 2012.

BOBADILLA, J.; ORTEGA, F.; HERNANDO, A.; GUTIERREZ, A.. **Recommender systems survey**. Knowledge-Based Systems , v. 46, p. 109-132, 2013.

ZHENG Y.; MOBASHER, Bamshad; BURKE, R.. CSLIM : **Contextual SLIM Recommendation Algorithms**. In Proceddings of the RecSys 2010 , p. 301-304, 2014.

CHENG, X.; DALE, C.; LIU J. Liu, J.. **Statistics and social network of Youtube videos**. In Procceddings of the 16th International Workshop on Quality of Service, p. 229-238, 2008.

CHUI, C. K.. **An Introduction to Wavelets**. San Diego:Academic Press, 1992.

DAVIDSON, J.; LIVINGSTON, B.; SAMPATH, D.; LIEBALD, B.;

LIU, J.; NANDY, P.; VAN VLEET, T.; GARGI, U.; GUPTA, S.; HE, Y.; LAMBERT, M.. **The YouTube video recommendation system**. In the Proceedings of the Fourth ACM conference on Recommender systems - RecSys '10 , p. 293, 2010.

DENG Z.; YAN, M.; SANG, J.; XU, C.. **Twitter is faster: Personalized time-aware video recommendation from twitter to Youtube**. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) , v. 11, n. 2, 2015.

KIM H. N.; JI, A. T.; HA, I.; JO, G.-S. (2010). **Collaborative Filtering based on collaborative tagging for enhancing the quality of recommendation**. Electronic Commerce Research and Applications , v. 9 n. 1, p. 73-83, 2010.