MODELAGEM, CARACTERIZAÇÃO E RECOMENDAÇÃO EM SERVIÇOS DE CONTEÚDO *WEB* MULTIMÍDIA

DIEGO DE MOURA DUARTE

MODELAGEM, CARACTERIZAÇÃO E RECOMENDAÇÃO EM SERVIÇOS DE CONTEÚDO *WEB* MULTIMÍDIA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais – Departamento de Ciência da Computação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Adriano César Machado Pereira
Coorientador: Clodoveu Augusto Davis Júnior

Belo Horizonte
Agosto de 2013

© 2013, Diego de Moura Duarte.
Todos os direitos reservados.

Duarte, Diego de Moura

D812m Modelagem, Caracterização e Recomendação em Serviços de Conteúdo *Web* Multimídia / Diego de Moura Duarte. — Belo Horizonte, 2013
xxiv, 58 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Orientador: Adriano César Machado Pereira
Coorientador: Clodoveu Augusto Davis Júnior

1. Computação – Teses. 2. Sistemas multimídia – Teses. 3. Modelagem de dados – Teses. 4. Sistemas de recomendação. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*75 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Modelagem, caracterização e recomendação em serviços de conteúdo web multimídia

DIEGO DE MOURA DUARTE

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ADRIANO CÉSAR MACHADO PEREIRA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. FLÁVIO LUIS CARDEAL PÁDUA
Departamento de Computação - CEFET/MG

PROFA. MIRELLA MOURA MORO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 28 de agosto de 2013.

Dedico este trabalho à minha família, companheira, e ao meu pai acadêmico.

Agradecimentos

Agradeço inicialmente a Deus, pela sabedoria e força ofertados.

Agradeço à minha mãe, pai, irmão e companheira, pelo apoio nos momentos mais difíceis, pela compreensão durante todos esses anos ausentes, e pelo amor incondicional oferecido todos os dias.

Agradeço à minha família e amigos, também pela compreensão durante os anos ausentes, e pelos momentos alegres existentes nos intervalos dos estudos.

Agradeço à Universidade Federal de Minas Gerais e Departamento de Ciência da Computação, pela oportunidade oferecida.

Agradeço à empresa Samba Tech, pelo tempo e recursos disponibilizados.

E agradeço principalmente ao meu pai acadêmico, pela motivação e confiança prodigalizados a todo momento, pelo exemplo de vida, e pela amizade durante todos esse anos.

"Digo: o real não está na saída nem na chegada:
ele se dispõe para a gente é no meio da travessia."
(João Guimarães Rosa)

Resumo

Nos últimos anos, a quantidade de conteúdo *Web* multimídia tem crescido significantemente. Um grande exemplo desse conteúdo é o vídeo online, como demonstrado pelo sucesso de plataformas como o YouTube. Esse crescimento também é observado em cenários corporativos, como emissoras de TV. Este trabalho apresenta um estudo de serviços de conteúdo *Web* multimídia em redes corporativas. Utilizando dados reais oferecidos pela Samba Tech, maior plataforma de distribuição de vídeos online da América Latina, propomos uma modelagem e caracterização desse tipo de serviço, assim como uma técnica de recomendação com foco no objeto sendo consumido. Resultados experimentais indicaram que o método proposto é muito promissor, chegando a quase 70% de precisão. Realizamos também análises distintas utilizando diferentes abordagens da literatura, como uma técnica estado-da-arte de recomendação de itens. Os resultados de nossa pesquisa são de grande importância para os provedores de conteúdo e seus consumidores, com aplicabilidade em serviços de personalização e sistemas de recomendação.

Palavras-chave: Vídeo Online, Conteúdo Multimídia, Modelagem, Caracterização, Recomendação.

Abstract

Web multimedia content has reached much importance lately. One of the most important content types is online video, as demonstrated by the success of platforms such as YouTube. The growth in the volume of available online video is also observed in corporate scenarios, such as TV stations. This work presents an analysis of corporate multimedia Web content services. We evaluate real data from online videos hosted by Samba Tech, the largest platform for online multimedia content distribution in Latin America. After modeling and characterize this service, we propose a novel technique for multimedia content recommendation, focusing on object being consumed. Experimental results indicate that the proposed method is very promising, obtaining almost 70% in precision. We also perform distinct evaluations using different approaches from literature, such as the state-of-the-art technique for item recommendation. These results are important for content providers and consumers, and it can be applied in personalized services and recommendation systems.

Keywords: Online Video, Multimedia Content, Modeling, Characterization, Recommendation.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

cCDF

Complementary Cumulative Distribution Function

CDN

Content Delivery Network

CF

Collaborative Filtering

DRM

Digital Rights Management

MAE

Mean Absolute Error

MIT

Massachusetts Institute of Technology

nDCG

Normalized Discounted Cumulative Gain

RMSE

Root Mean Squared Error

SaaS

Software as a Service

STTM

Samba Tech Tracking Module

UGC

User Generated Content

URL

Uniform Resource Locator

UTM

Urchin Tracking Module

WRMF

Weighted Regularized Matrix Factorization

Sumário

Capítulo 1

Introdução

Nos últimos anos a interação entre os usuários e a *Web* tem passado por grandes alterações. Com o advento da *Web 2.0*, o usuário deixou de ser um mero espectador que consume informações, passando a ser também provedor de conteúdo, o que resultou em uma crescente quantidade de informação disponível. Um dos exemplos mais significativos desse crescimento é o conteúdo multimídia, impulsionado pelo grande número de *gadgets* ou aparelhos eletrônicos, como celulares, câmeras e tocadores de música, vendidos atualmente.

Um representante significativo desse conjunto de multimídia é o vídeo online, que recebeu grande atenção após o surgimento e popularização do Youtube, maior serviço de conteúdo multimídia da atualidade. Sendo conhecido atualmente em uma escala mundial, suas estatísticas de uso impressionam[1]: são enviados atualmente 100 horas de vídeos por minuto, ocorrem mais de 6 bilhões de visualizações mensais, e está localizado em 56 países, sendo distribuído em 61 idiomas.

Na maioria dos casos de serviços de conteúdo multimídia, utiliza-se o modelo UGC (*User Generated Content*), onde o usuário é responsável por gerar e consumir tal conteúdo. Exemplos desse serviço, além do YouTube, são: Vímeo, Flickr e SoundCloud. Porém, a obtenção de seus dados é um obstáculo, e muitas vezes esse processo é realizado através de *bots*[2] ou *crawlers*[3], o que gera informações de pouca confiança e cobertura restrita.

Esta dissertação investiga o cenário de conteúdo multimídia disponibilizado na *Web* com o objetivo de modelar, caracterizar e aperfeiçoar esses serviços multimídia

[1]Fonte: *YouTube Statistics* (`http://www.youtube.com/t/press_statistics`). Online; Acessado em 01-Janeiro-2011

[2]Aplicação de *software* concebida para simular ações humanas repetidas vezes de maneira padrão.

[3]Também conhecido como *Internet bot*, navega sistematicamente na *World Wide Web*, tipicamente com o propósito de indexação da *Web*.

em um contexto corporativo, trazendo vantagens para o provedor e consumidor do conteúdo. No restante deste capítulo, apresentamos a motivação do trabalho na Seção 1.1. Em seguida, na Seção 1.2, são descritos os principais objetivos, e apresentadas as contribuições na Seção 1.3. Por fim, sintetizamos a organização dessa dissertação de mestrado (Seção 1.4).

1.1 Motivação

Observa-se também uma popularização crescente do uso de conteúdo *Web* multimídia no âmbito corporativo. Grandes emissoras de TV, nacionais e internacionais, estão percebendo essa mudança de comportamento dos usuários, e muitas já disponibilizam grande parte de seu conteúdo online. Um grande exemplo é a Rede Globo, maior rede de televisão nacional, que exibe em seu site praticamente toda a sua programação.

Diferentemente do modelo *UGC*, nesse cenário o provedor do conteúdo é a emissora. Consequentemente, o conteúdo disponibilizado é mais restrito, além dos dados necessários para o seu estudo serem de difícil obtenção, o que é evidenciado pela carência de estudos investigativos nesse contexto.

Através de um protocolo de cooperação, obtivemos uma base de dados real e representativa da Samba Tech [Sam04], empresa que possui a maior plataforma de vídeos online da América Latina. Dentre os seus maiores clientes, podemos citar: SBT (Sistema Brasileiro de Televisão), iG (*Internet Group*), Rede Bandeirantes, Anhanguera Educacional e *El Comercio*.

1.2 Objetivos

Os objetivos gerais deste trabalho são:

> *Modelar, caracterizar e aperfeiçoar serviços* Web *multimídia, com enfoque em recomendação de conteúdo.*

Como objetivos específicos desta dissertação, estão:

- **Modelar** o serviço *Web* Multimídia, com o intuito de obter uma representação adequada de seu funcionamento;
- **Caracterizar** esse serviço para um melhor entendimento de suas características e especificidades;

- Propor, implementar e validar um novo **modelo de recomendação** de conteúdo multimídia;
- **Avaliar** a qualidade da técnica de recomendação, desenvolvida com base neste modelo, em comparação com outros métodos correlatos da literatura.

1.3 Contribuições do trabalho

As principais contribuições do trabalho apresentado nessa dissertação são as seguintes:

- Um melhor entendimento do funcionamento acerca de serviços de conteúdo multimídia na *Web*, em um âmbito corporativo;
- Uma visão geral de aspectos quantitativos desse tipo de aplicação *Web*;
- Um novo modelo de recomendação com base no consumo de objetos multimídia;
- A implementação de uma técnica de recomendação baseada nesse modelo;
- Um método de validação da recomendação de conteúdo multimídia;
- A avaliação do modelo de recomendação proposto, utilizando tal método de validação, aplicado em dados reais.

Os resultados de nossa pesquisa também possuem grande importância para os provedores de conteúdo e seus consumidores. A partir do entendimento do funcionamento de tais serviços, é possível propor diversas melhorias que serão relevantes para ambas as partes. Como exemplo, citam-se a personalização de serviços, além da própria técnica de recomendação desenvolvida. Com um foco no consumo de objeto, tal técnica pode ser de grande utilidade para a melhoria de serviços de conteúdo multimídia em geral. Os resultados apresentados neste trabalho apresentam uma avaliação do impacto dessa técnica utilizando dados reais, o que, por sua vez, demostram a sua qualidade e eficácia.

1.4 Organização da dissertação

Esta dissertação está organizada da seguinte maneira: o Capítulo 2 contém os trabalhos correlatos com a nossa pesquisa. O Capítulo 3 apresenta a modelagem representativa de serviços de conteúdo multimídia. Realizamos uma caracterização de dados do serviço da plataforma Samba Tech no Capítulo 4. No Capítulo 5, temos a proposta de um

modelo e técnica de recomendação para esse tipo de serviço. As análises e resultados da aplicação dessa técnica são expostos no Capítulo 6. Por fim, apresentamos as conclusões e trabalhos futuros no Capítulo 7.

Essa organização pode ser melhor visualizada através do Mapa Mental apresentado na Figura 1.1

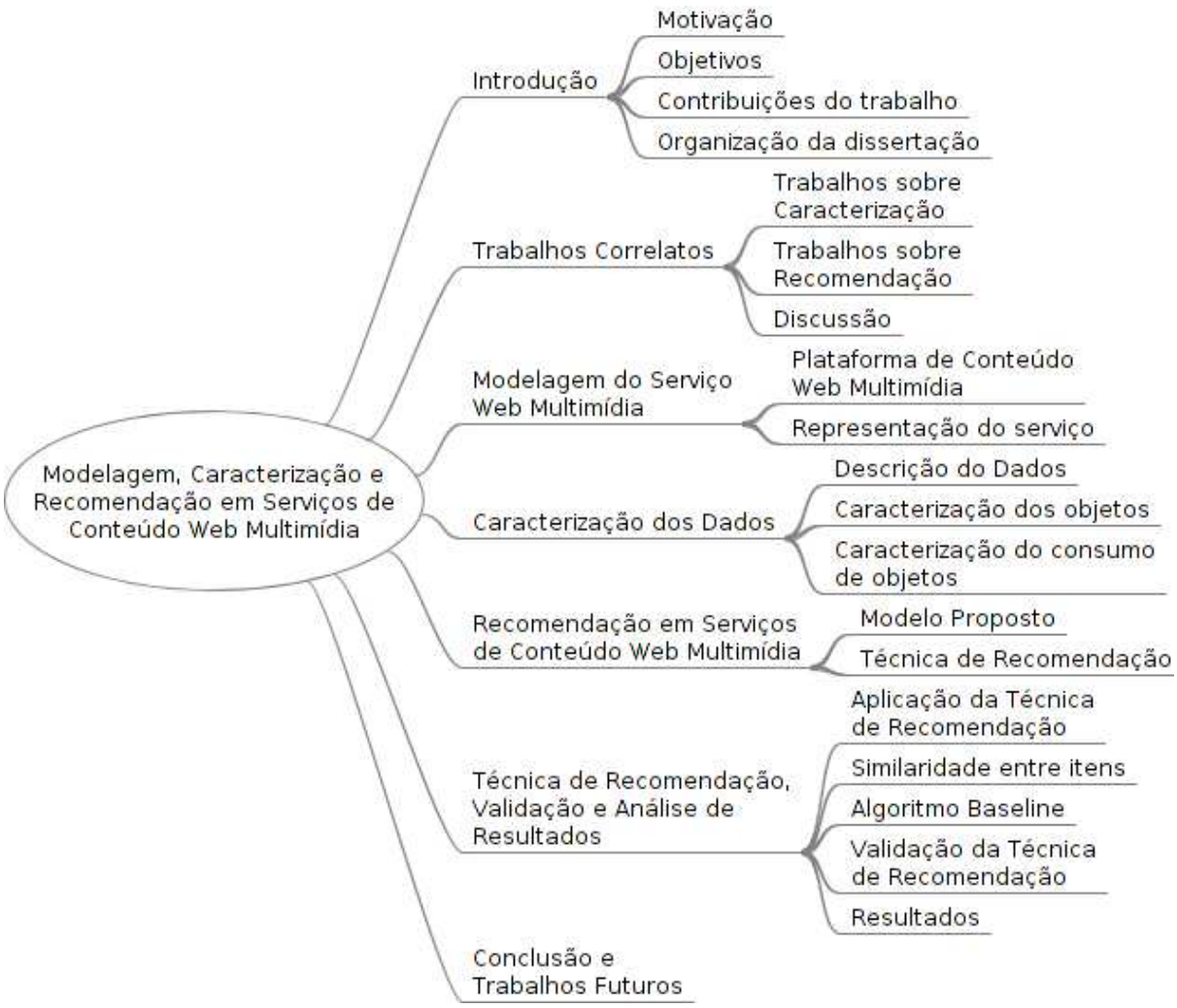


Figura 1.1. Mapa mental - Estrutura da Dissertação.

Capítulo 2

Trabalhos Correlatos

A seguir, são apresentados trabalhos relacionados com a pesquisa desenvolvida nesta dissertação. Separamos os trabalhos com relação aos principais tópicos abordados: na Seção 2.1, listamos os trabalhos relacionados à caracterização de dados, e em seguida, os trabalhos que abordam o assunto de recomendação, Seção 2.2. Na Seção 2.3, temos uma discussão sobre a relação dessas referências com a pesquisa desenvolvida nessa dissertação.

2.1 Trabalhos sobre Caracterização

Cha *et al.* [CKR+09] utilizam em seu trabalho os serviços do YouTube e Daum Videos (Coréia) para demonstrar empiricamente como serviços do tipo UGC (*User Generated Content* são fundamentalmente diferentes de serviços tradicionais de vídeo sob demanda. Também é realizada uma avaliação da popularidade dos vídeos, onde observou-se uma distribuição exponencial e demonstrou-se, assumindo uma distribuição de Zipf [Zip32] como base, que é possível aumentar em 45% o número de visualizações ao remover gargalos. Estudando a evolução dos vídeos ao longo do tempo, os autores demonstraram que a popularidade de um vídeo é concentrada no período imediatamente após sua publicação.

Cheng *et al.* [CLD13] realizam um trabalho de análise dos vídeos do YouTube, obtidos a partir de um *crawler* e chegando a um número de mais de 2,5 milhões de vídeos. Seus autores avaliam algumas características, como popularidade das categorias e número de visualizações. Além disso, eles investigam a rede social dos vídeos do YouTube, criada pelos vídeos relacionados e conteúdo gerado pelo usuário. Tal rede possui características de *"small-world"* e de coeficiente elevado de agrupamento (*clustering*),

o que significa que esse comportamento pode ser explorado para facilitar o design de *caching* e estratégias *"peer-to-peer"* para compartilhamento de vídeos.

Os trabalhos de Acharya *et al.* [ASP00] e Chesire *et al.* [CWVL01] tiveram como foco o estudo da popularidade. O primeiro realizou suas análises com base em acessos de usuários a vídeos transmitidos na *Web*, e identificou que a popularidade do conteúdo não segue uma distribuição Zipf. Já o segundo analisou a carga de trabalho de um servidor de fluxo de mídias de uma grande empresa, e observou que a sua popularidade segue uma distribuição de Zipf. Ambos os trabalhos chegam a resultados distintos, o que pode ser explicado pela diferença entre os conteúdos avaliados, ressaltando o impacto do ambiente de estudo.

Em Veloso *et al.* [VAM+02], os autores estudaram a carga de trabalho de um servidor comercial de fluxos de vídeo ao vivo localizado no Brasil. O foco do estudo foi a caracterização do processo de chegada de sessões de usuários, bem como dos tempos dessas sessões, com o intuito de utilizar os resultados em um gerador de cargas sintéticas realistas. Os autores consideraram um modelo composto por períodos de atividade (quando o usuário recebe mídia) intercalados por períodos de inatividade, disparados por alguma ação do usuário (como pausa). Alguns dos principais resultados obtidos são: (1) o processo de chegada dos usuários segue uma distribuição de Poisson; (2) os períodos de atividade são bem modelados por uma distribuição log-normal; (3) os períodos de inatividade são bem modelados por uma distribuição exponencial.

Sripanidkulchai *et al.* [SMZ04] realizaram uma caracterização de uma carga de trabalho de fluxos de vídeo e áudio ao vivo de uma grande CDN (*Content Delivery Network*), o que possibilitou a análise de uma ampla diversidade de conteúdos. Mais de 90% do conteúdo analisado foi apenas áudio. Observou-se que a popularidade dos conteúdos segue uma distribuição Zipf de duas partes. Foi também observado que os clientes entram no sistema de acordo com uma distribuição exponencial e que a duração das sessões dos usuários apresenta cauda pesada.

O trabalho de Benevenuto *et al.* [BPR+09] apresenta uma análise da carga de trabalho de um serviço de vídeos do Universo OnLine (UOL), utilizando sessões e requisições ao servidor. Realiza também uma avaliação do perfil de navegação dos usuários desse serviço. Através da modelagem dos padrões dessa navegação, os autores identificaram grupos de usuários com diferentes padrões de acesso, o que pode ser útil para prover novas políticas de personalização ou recomendação para usuários.

No trabalho de García *et al.* [GPnMG09], é proposto um modelo probabilístico mais realístico para melhorar a predição de cargas de trabalho em simulações e ambientes de teste. Para isso, é realizada uma caracterização das interações do usuário em um serviço de vídeo sobre demanda, onde são avaliadas ações como *play*, *pause* e

stop relacionadas a um determinado vídeo ou conjunto desses. Por ser um fator comum, os autores utilizam o tamanho do vídeo como elemento base, juntamente com distribuições univariadas, para determinar características de interações de usuário.

Em 2011, Gonçalves [GTD+11] propôs uma metodologia de caracterização hierárquica do conteúdo multimídia organizada de forma hierárquica em quatro partes: Requisição (***R****equest*), Objeto (***O****bject*), Conteúdo (***C****ontent*) e Conhecimento (***K****nowledge*) - ROCK. A metodologia propõe uma segmentação das análises em diferentes camadas visando a extração de informações existentes nos conteúdos. Além disso, o autor aplica essa metodologia em dados reais de um distribuidor de conteúdos multimídia corporativo, demonstrando sua utilidade e aplicabilidade.

2.2 Trabalhos sobre Recomendação

Definir recomendação, caso não seja definido no futuro...

Com relação à recomendação, existe uma grande quantidade de trabalhos que possuem o objetivo de apresentar técnicas para sistemas de recomendação. Em 2011, foi publicado o *Recommender Systems Handbook* [RRSK11], que apresenta um conjunto de artigos que envolvem os seguinte tópicos: técnicas, aplicações e avaliações de sistemas de recomendação; interações com sistemas de recomendação; sistemas de recomendação e comunidades; e algoritmos avançados. Este estudo aborda assuntos que compõem a base de sistemas de recomendação, sendo utilizado como referência para a aplicação e desenvolvimento do método de recomendação utilizado em nossa pesquisa.

Su e Khoshgoftaar [SK09] apresentam inúmeras técnicas de *Collaborative Filtering* (CF), uma das abordagens mais bem sucedidas para construção de sistemas de recomendação. A partir de uma descrição de suas principais vantagens e desvantagens, são descritas as principais técnicas de CF: *memory-based*, *model-based* e híbridas (combinação das duas primeiras).

Um dos atuais desafios refere-se à modelagem do comportamento do usuário. Sistemas de recomendação são baseados em perfis [JZFF11]. O conhecimento de preferências e interesses de cada usuário é importante para a identificação de potenciais itens relevantes. No entanto, cada usuário pode ser modelado através de conjuntos distintos de objetos, o que torna a definição do melhor modelo uma tarefa complexa. Alguns estudos recentes [LK04] demonstram a necessidade da definição de perfis mais amplos e informativos.

Outros desafios existentes são referentes a dados esparsos e *cold start*. No primeiro, temos um número de objetos bem superior à quantidade de usuários, que conso-

mem apenas uma pequena porção desses itens. Além disso, existe uma grande concentração de usuários ao redor de poucos objetos distintos, seguido por uma concentração decrescente sobre outros objetos, gerando o fenômeno de *long tail* [And10]. O segundo refere-se à dificuldade de gerar recomendações de novos itens ou para novos usuários, uma vez que inicialmente existe pouca informação sobre eles [SPUP02].

Muitos trabalhos de recomendação utilizam como base o Youtube. Baluja *et al.* [BSS+08] apresentam uma técnica com base em grafos construídos a partir do histórico de visualizações do usuário. Davidson *et al.* [DLL+10] descreveram o sistema de recomendação do Youtube com foco em técnicas do tipo *Top-N*, levando em consideração o conteúdo do vídeo (e.g., metadados) e as interações do usuário (e.g., classificação de vídeos) para a criação dos *rankings*. Sua classificação possui como base sinais (características dos vídeos, histórico do usuário, etc.), que são combinados linearmente para a geração de *rankings*, o que resulta em uma recomendação de 4 a 60 objetos.

2.3 Discussão

As referências apresentadas na Seção 2.1 são importantes para o direcionamento da caracterização realizada em nossa pesquisa. Podemos destacar os trabalhos de Benevenuto [BPR+09] e Gonçalves [GTD+11]. O primeiro contribui para o entendimento do conceito de sessões de usuários e de seus padrões de navegação. O segundo oferece um aprendizado referente ao consumo de conteúdo multimídia, propondo uma metodologia que pode ser aplicada em nosso trabalho, segmentando nossas análises de acordo com os perfis de acesso identificados. Sendo assim, podemos explorar a camada de conhecimento definida nessa metodologia a partir de nossas análises e base de dados, o que contribui para a identificação dos padrões de acessos a conteúdos multimídia.

No campo de recomendação na Seção 2.2, as referências exploram diversas métodos de recomendação, assim como inúmeras técnicas aplicáveis nesse contexto. Similarmente à nossa pesquisa, o trabalho de Davidson *et al.* [DLL+10] apresenta uma técnica de recomendação baseada na geração de *rankings*. Esta dissertação se distingue das referências citadas pelo cenário da aplicação utilizado, composto por dados corporativos. Além disso, apresenta um foco maior no consumo de objetos, desconsiderando o perfil do usuário que o acessa.

Todos os trabalhos correlatos contribuem com ideias e técnicas para auxiliar na modelagem do problema de caracterização e recomendação de conteúdo multimídia endereçado nesta pesquisa. Nossa abordagem, ao propor o uso de modelos já consolidados na literatura com uma nova visão focada no consumo do objeto, define um novo po-

tencial de ganho e aplicação a diferentes cenários da *Web*, que demandam mecanismos cada vez mais robustos e personalizados para recomendação.

Capítulo 3

Modelagem do Serviço *Web* Multimídia

Neste capítulo, apresentamos uma modelagem de serviços de conteúdo *Web* multimídia. Na Seção 3.1 descrevemos a plataforma utilizada como base e, em seguida, demonstramos um modelo para a representação desse tipo de serviço na Seção 3.2.

3.1 Plataforma de conteúdo *Web* multimídia

Atualmente, existem diversos serviços que oferecem uma plataforma de conteúdo *Web* multimídia. Dentre eles, citam-se serviços focados em áudio, como o SoundCloud; em imagem, como o Flickr; e em vídeos, como o Vímeo e o Youtube. Grande parte desses serviços possuem o modelo de UGC (*User Generated Content*), onde o usuário é responsável por gerar e consumir tal conteúdo. Dessa forma, eles oferecem meios para o envio, armazenamento, gerenciamento e acesso a esse conteúdo. Este mesmo cenário é encontrado no ambiente corporativo, com a exceção de que o provedor e gestor do conteúdo não é o usuário, mas sim a empresa em questão.

A seguir, a Seção 3.1.1 apresenta a plataforma corporativa da Samba Tech. Em seguida, a Seção 3.1.2 descreve uma de suas mais importantes ferramentas: o Samba Tech Tracking Module.

3.1.1 Samba Tech

A Samba Tech [Sam04] é uma empresa nacional fundada em 2004, que atua no mercado de *Software as a Service* (SaaS) oferendo soluções de gerenciamento e distribuição de conteúdo multimídia na Internet. Possuindo como foco principal os vídeos online,

possui uma plataforma para esse tipo de conteúdo considerada a maior da América Latina. A empresa suporta um tráfego anual de 14PB e garante a distribuição de quase meio milhão de mídias (vídeo, imagem ou áudio). Recebe investimentos da DFJ FIR Capital e tem parceria global com o MIT (*Massachusetts Institute of Technology*). Seus maiores clientes incluem grandes emissoras de televisão, como SBT e Rede Bandeirantes, portais de notícia como *El Comercio*, grupos como Abril e iG, além de redes de educação, como a Anhanguera.

Os serviços oferecidos pela Samba Tech incluem um sistema *online* para o envio e gerenciamento de conteúdo multimídia, com um armazenamento transparente para o seu provedor. Tal conteúdo pode ser um vídeo, imagem ou áudio. Outras funcionalidades incluem o *encoding* de vídeos, segurança de conteúdo a partir de seu domínio ou geolocalização, encriptação de vídeos para solução de DRM (*Digital Rights Management*), monetização de vídeos, entrega de conteúdo através de serviços de CDN (*Content Delivery Network*), entre outros.

Uma ferramenta importante é o *player* (Figura 3.1), desenvolvido pela Samba Tech, que é responsável por servir tais conteúdos em diversos ambientes, como diferentes modelos e versões de navegadores (*browsers*) ou *devices* (e.g., celulares, *tablets*, etc.). Sendo a única forma de interação entre o usuário e o provedor de conteúdo, é possível realizar diversas customizações dessa ferramenta, como mudança de suas cores e dimensões.

Figura 3.1. *Player* da Samba Tech.

Por fim, uma das funções mais importantes do sistema oferecido pela Samba Tech é a coleta de dados para análise estatística. Para essa finalidade, foi-se desenvolvida recentemente a ferramenta Samba Tech Tracking Module, detalhada na seção seguinte. Um sumário do gama de serviços oferecidos por essa empresa pode ser visto

na Figura 3.2

Figura 3.2. Serviços oferecidos pela Samba Tech.

3.1.2 Samba Tech Tracking Module

O *Samba Tech Tracking Module* (STTM), é um modelo para rastreamento de diversas métricas na *Web*, assim como a visualização de informações e estatísticas destes dados. Tal modelo é inspirado no *Urchin Tracking Module* (UTM), um programa de análise de estatísticas da *Web* desenvolvido pela *Urchin Software Corporation* para análise de arquivos de *log* de servidores e disponibilização de informações de seu tráfego.

O STTM foi desenvolvido pela Samba Tech, contando com grande participação do autor dessa dissertação. A partir de uma integração com o *player* da plataforma da empresa, é possível realizar a coleta de diversos eventos ocorridos nesse contexto, como o carregamento (*load*) de um *player* em uma página, assim como os eventos de *play* e *pause*, dentre outros. A partir dessa coleta, podemos obter inúmeras informações, como, por exemplo, o número de visualizações de um determinado vídeo, assim como a quantidade relativa que um usuário assistiu de uma mídia, sendo 100% a visualização completa da duração do vídeo.

Esta dissertação utiliza os dados coletados pelo STTM. A partir dessa coleta, é possível realizar a caracterização, identificação de padrões, e propor melhorias para o serviço, como demonstrado nas próximas seções.

3.2 Representação do serviço

Com base na plataforma de conteúdo *Web* multimídia descrita na seção anterior, podemos montar um modelo para a representação desse tipo de serviço. Esse processo é de crucial importância para o entendimento do problema abordado nesta pesquisa, e serve como base para o estudo realizado no restante do trabalho.

Inicialmente, é necessário definir o contexto de desenvolvimento deste trabalho. Diferentemente dos serviços que utilizam o modelo de *User Generated Content*, nossa pesquisa possui como foco o ambiente corporativo, onde os provedores de conteúdo não são os usuários, mas sim empresas produtoras de conteúdo. Dessa forma, nossa modelagem define dois conjuntos de entidades:

- **Provedores de conteúdo**: formado por empresas produtoras de conteúdo multimídia, como grandes emissoras de televisão, instituições educacionais, dentre outras. Suas responsabilidades envolvem o envio e gerenciamento do conteúdo multimídia, assim como o meio para a sua exibição, que pode ser realizado através de *Web sites*, portais online, ou até mesmo páginas do Facebook.

- **Consumidores**: formado por usuários comuns, possuem como única responsabilidade o acesso a conteúdos multimídia.

Uma plataforma para o serviço em questão deve oferecer meios para que ambas as entidades possam realizar o processo de envio, gerenciamento e exibição de conteúdo multimídia. Dessa forma, são responsabilidades desse serviço:

1. **Envio (*Upload*)**: inicialmente, é necessário oferecer meios de envio de vídeos, áudio, imagens, etc.;

2. **Armazenamento**: após o envio de um arquivo, é responsabilidade da plataforma o seu armazenamento, o que deve ocorrer de maneira transparente para o seu produtor;

3. **Gerenciamento**: finalizado o envio do conteúdo, o seu gerenciamento também é uma funcionalidade a ser oferecida pela plataforma. Esta tarefa inclui a recuperação do conteúdo enviado, assim como a edição de seus metadados (e.g., nome, título, descrição);

4. **Acessibilidade**: por fim, a plataforma deve oferecer recursos para que o conteúdo inicialmente enviado seja acessado por usuários.

Dessa forma, a Figura 3.3 ilustra um modelo representativo para um serviço de conteúdo *Web* multimídia.

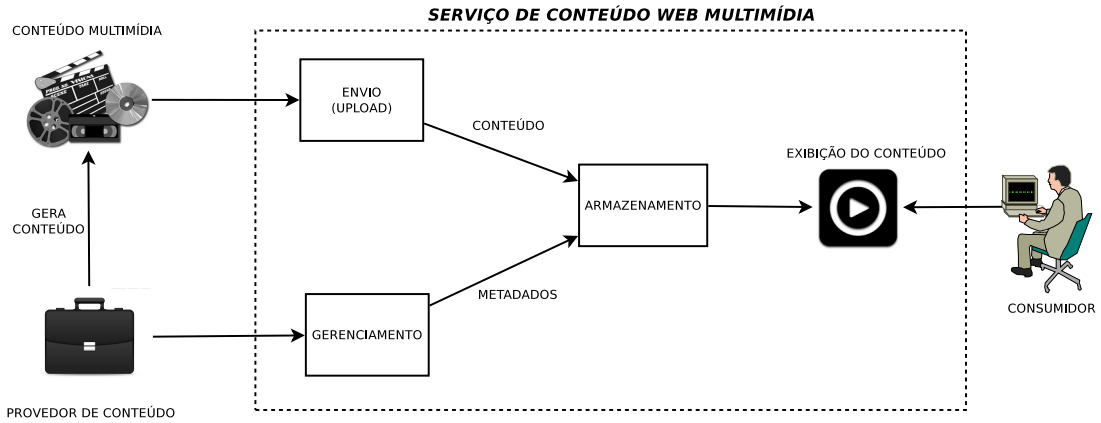


Figura 3.3. Modelo representativo de serviços de conteúdo *Web* multimídia.

Com base na representação apresentada, podemos modelar a plataforma da Samba Tech na Figura 3.4. Essa figura ilustra a ferramenta de gerenciamento de conteúdo da Samba Tech, chamada de Samba Vídeos. Também são apresentados alguns dos serviços utilizados para o armazenamento de conteúdo e metadados: *Amazon Web Services* e o gerenciador de bancos de dados *MySQL*. Por fim, indicamos a entrega do conteúdo através de um serviço de CDN, exemplificado pela *EdgeCast*.

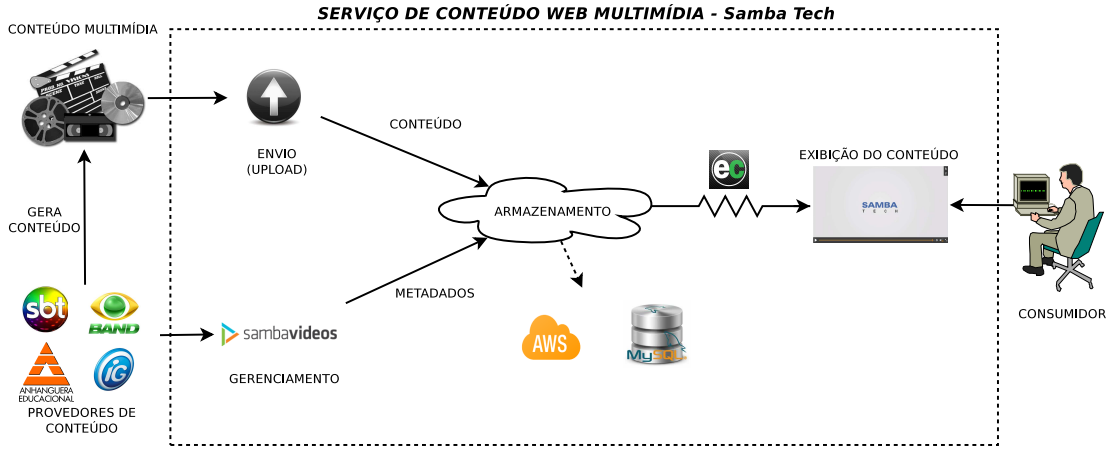


Figura 3.4. Modelo representativo da Samba Tech.

Após a modelagem do serviço de conteúdo *Web* multimídia da Samba Tech, apresentamos no capítulo seguinte a descrição e caracterização de seus dados.

Capítulo 4

Caracterização dos Dados

Neste capítulo, apresentamos a caracterização dos dados utilizados em nossa pesquisa. A Seção 4.1 apresenta a descrição desses dados. As duas seções seguintes descrevem uma caracterização dos dados em dois conjuntos: o primeiro, Seção 4.2, com a caracterização com foco nos objetos, e o segundo, Seção 4.3, com o foco no consumo desses objetos.

4.1 Descrição dos Dados

Os dados obtidos para este estudo foram obtidos a partir do Samba Tech Tracking Module, ou STTM, modelo detalhado na Seção 3.1.2. Esses dados são coletados a partir do *player* fornecido pela plataforma da Samba Tech e, consequentemente, possuem informações de todas as visualizações de usuários.

Inicialmente, devemos compreender o formato dos dados coletados pelo STTM. Para esse propósito, o modelo definiu o conceito de **sessão**. No cenário estudado, similar ao conceito definido no trabalho de Veloso [VAM+02], uma sessão é composta por todas as interações realizadas por um usuário (visualização, *play*, *stop*, *resume*, etc.) em um determinado *player*. Então, uma sessão está vinculada a um único objeto sendo consumido, e todas as interações do usuário estão presentes em uma mesma sessão.

Além dos dados obtidos a partir do STTM, a empresa Samba Tech forneceu grande parte de seu banco de dados. Devido à arquitetura de sua plataforma, inúmeros metadados de seus conteúdos estão armazenados apenas em seu banco de dados. Dessa forma, com a sua obtenção, podemos realizar um estudo de maior qualidade.

Os dados coletados, considerando a ferramenta STTM e a base de dados, contemplam todas as sessões ocorridas entre o período de **01 de Julho de 2012** a **31**

de Julho de 2012, que totalizam aproximadamente 60 milhões de sessões em quase 80GB de arquivos (*logs*). Além disso, essa plataforma possui basicamente três tipos de conteúdo multimídia: vídeo, áudio e imagem. A Tabela 4.1 apresenta a distribuição desse conteúdo na base de dados da plataforma em questão.

Conteúdo multimídia	Ocorrência
Vídeo	92,40%
Imagem	4,57%
Áudio	2,28%
Outros[1]	0,75%

Tabela 4.1. Distribuição de conteúdo multimídia na plataforma Samba Tech.

De acordo com os dados expostos na Tabela 4.1, todos os estudos a seguir são realizados considerando-se apenas vídeos como objetos, já que representam quase a totalidade da base de dados da empresa Samba Tech.

Nas seções seguintes, serão caracterizados os dados coletados. Esse processo foi realizado com a distinção de dois grupos: objeto e seu consumo. O primeiro é um estudo com foco nos metadados do objetos, enquanto o segundo aborda informações do seu consumo, com o estudo do popularidade dos vídeos da plataforma Samba Tech.

4.2 Caracterização dos objetos

A seguir, demonstramos algumas caracterizações realizadas sobre os objetos e seus metadados. Nesse processo, é realizada uma comparação entre os objetos do STTM e os objetos da base de dados. Essa distinção é realizada porque nem todos os vídeos da base de dados da plataforma Samba Tech foram visualizados. Dessa forma, a diferença entre esses dois grupos é que o primeiro (STTM) corresponde ao conjunto de vídeos que tiveram pelo menos uma sessão coletada pelo Samba Tech Tracking Module durante o mês estudado, enquanto o segundo grupo (Base de dados) corresponde a todos os objetos da base de dados da Samba Tech.

4.2.1 Tempo de duração

A análise desta seção envolve o tempo de duração de todos os vídeos de nossa base de dados. A Tabela 4.2 apresenta algumas estatísticas sobre essa duração. Para a

[1]Incluem mídias com tipo de conteúdo indefinido

análise de percentis, foram utilizados apenas os valores distintos da duração dos vídeos de nosso conjunto de dados, independente da quantidade de suas ocorrências.

	STTM	Base de Dados
Valores distintos de duração	154	216
Menor duração	0 minutos	0 minutos
Maior duração	234 minutos	3083 minutos
Percentil 25	37 minutos	53 minutos
Percentil 50	76 minutos	107 minutos
Percentil 75	114 minutos	161 minutos
Percentil 90	138 minutos	210 minutos
Percentil 99	197 minutos	1583 minutos

Tabela 4.2. Estatísticas do tempo de duração de vídeos.

A partir da Tabela 4.2, percebe-se que, na prática, vídeos muito grandes em termos de duração tendem a não são visualizados, mesmo que existam na base de dados. Por outro lado, de acordo com os percentis, vídeos extremamente longos são apenas exceções.

A cCDF (*Complementary Cumulative Distribution Function*) da duração para as mídias do STTM e da base de dados é exibida na Figura 4.1.

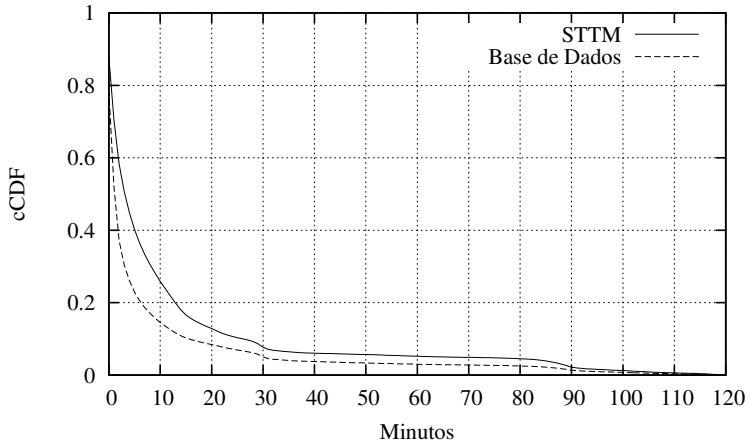


Figura 4.1. cCDF do tempo de duração dos vídeos.

Através da Figura 4.1, percebe-se que não existe muita distinção entre os grupos STTM e a base de dados. Também pode-se dizer que a grande maioria dos vídeos possuem duração abaixo de 10 minutos, já que correspondem a quase 80% dos dados.

4.2.2 Gênero

Esta seção caracteriza o gênero de todas as mídias. Essa informação é inserida pelo provedor de conteúdo durante a edição de metadados. Os possíveis valores são: Animais, Ciência, Comédia, Entretenimento, Esportes, Filmes, Música, Pessoas e Política. Além disso, é possível que o vídeo não possua nenhum gênero (opção: Sem gênero). A Figura 4.2 apresenta um histograma da distribuição desses gêneros entre os dois grupos avaliados.

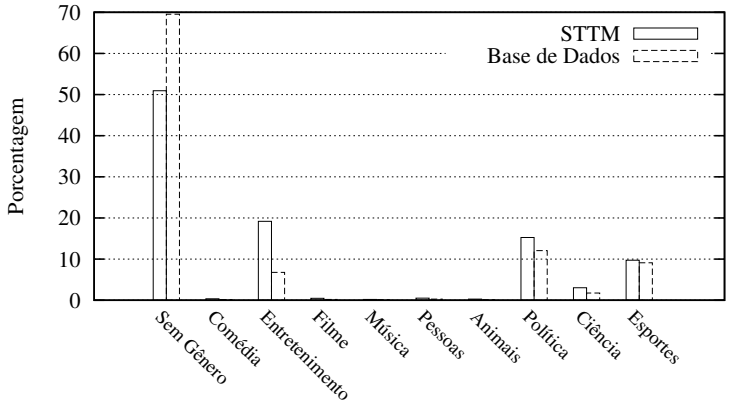


Figura 4.2. Histograma de distribuição de gêneros.

A Figura 4.2 mostra que a grande quantidade de mídias não possuem gênero na Base de Dados (cerca de 70%). Essa mesma superioridade se mantém para as mídias visualizadas pelo STTM, embora em uma proporção menor (cerca de 50%). Porém, em ambos os grupos, os gêneros que aparecem com a maior quantidade de mídias são entretenimento, política e esportes.

4.2.3 *Tags*

Os resultados expostos nessa seção se referem ao estudo de *tags*. Assim como o gênero, essa informação é inserida pelo provedor de conteúdo com diversos objetivos, como a simples classificação de mídias. Seu valor corresponde a uma lista de nomes ou *strings*. Algumas de suas estatísticas estão expostas na Tabela 4.3.

A Tabela 4.3 informa que existem no máximo 56 *tags* para uma mídia visualizada, enquanto este número chega a 90 para todas as mídias da base de dados. De acordo com os percentis, existe uma quantidade pequena de mídias com mais de 54 *tags* para a base de dados.

	STTM	Base de Dados
Valores distintos de quantidade de *tags*	53	59
Menor valor	0 *tags*	0 *tags*
Maior valor	56 *tags*	90 *tags*
Percentil 25	13 *tags*	14 *tags*
Percentil 50	26 *tags*	29 *tags*
Percentil 75	39 *tags*	44 *tags*
Percentil 90	47 *tags*	54 *tags*
Percentil 99	56 *tags*	90 *tags*

Tabela 4.3. Estatísticas de *tags* dos vídeos.

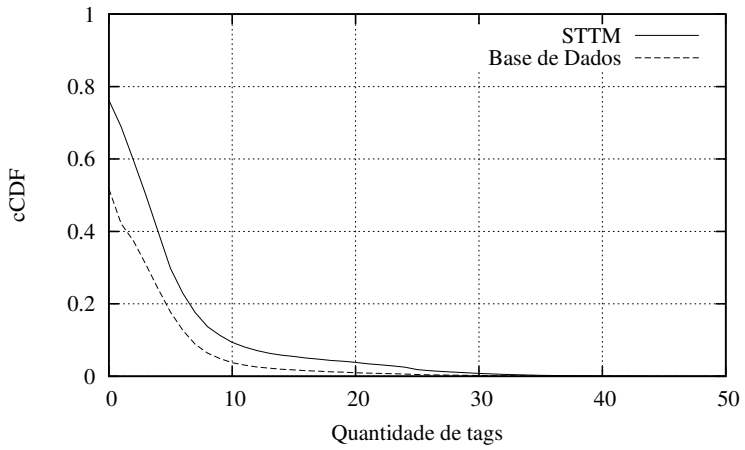


Figura 4.3. cCDF para quantidade de *tags* dos vídeos.

A Figura 4.3 demonstra que a grande maioria de mídias (cerca de 90%) possui menos de 10 *tags*. Porém, a quantidade de mídias com menos *tags* é maior para as mídias da base de dados do que para as mídias visualizadas.

4.3 Caracterização do consumo de objetos

Esta seção apresenta uma caracterização do consumo do objeto na plataforma da Samba Tech. Caracterizamos a popularidade do objeto no período do mês de Julho de 2012, assim como a distribuição de usuários distintos que consomem um objeto, a partir da premissa que nosso estudo apresenta foco no objeto sendo consumido.

A Figura 4.4 apresenta uma distribuição das visualizações ocorridas no *player* da Samba Tech agregadas pela hora do dia, sendo que este valor varia entre 0 (meia-noite) e 23 horas.

A partir da Figura 4.4, percebemos claramente a diferença de acessos entre a

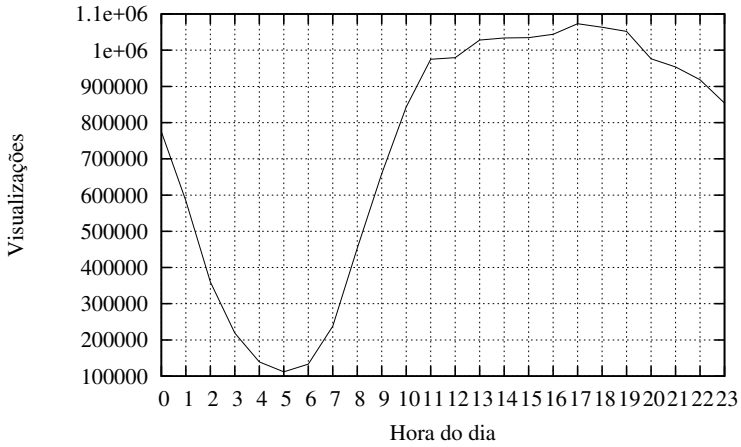


Figura 4.4. Distribuição de visualizações por hora.

madrugada e o restante do dia. A partir das 0 horas, ou meia noite, a quantidade de visualizações decresce drasticamente, chegando a um valor mínimo aproximadamente às 5 horas da madrugada. A partir desse horário, os acessos ao *player* começam a crescer e estabilizam por volta de 10 horas da manhã. O valor máximo é alcançado às 17 horas, com aproximadamente 1,07 milhões de visualizações. É importante notar que esse valor é agregado no mês de Julho, ou seja, a soma de todas as visualizações que ocorreram entre 17 e 18 horas durante 31 dias. Essa mesma agregação de visualizações é apresentada na Figura 4.5, porém agrupadas por dia.

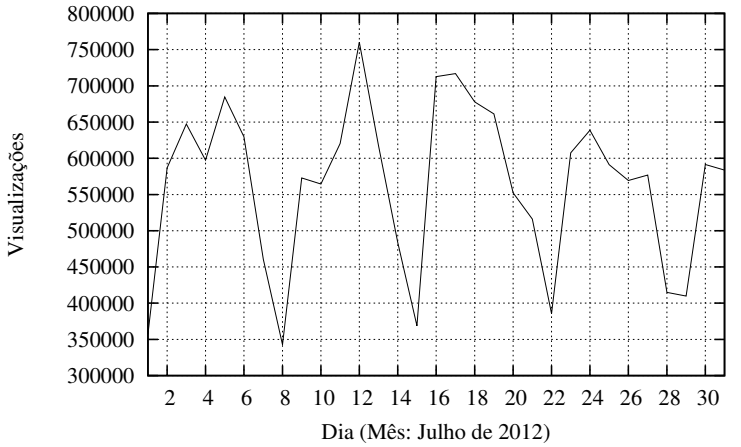


Figura 4.5. Distribuição de visualizações por dia.

A Figura 4.5 demonstra uma média de 600 mil visualizações por dia. Embora a quantidade de sessões fornecidas seja de 60 milhões, em muitos casos um *player* é carregado em uma página da *Web*, mas não ocorre de fato uma visualização (ausência do evento de *play*). Dessa forma, esse evento foi desconsiderado em todas as métri-

cas desse estudo, uma vez que não podemos dizer que o usuário assistiu à mídia. A Figura 4.5 também mostra que, durante os finais de semana, a quantidade de sessões decresce significativamente, enquanto durante a semana não é possível identificar um comportamento uniforme.

O último gráfico de agregação de visualização é exposto na Figura 4.6, onde os dias da semana são utilizados para o agrupamento de visualizações.

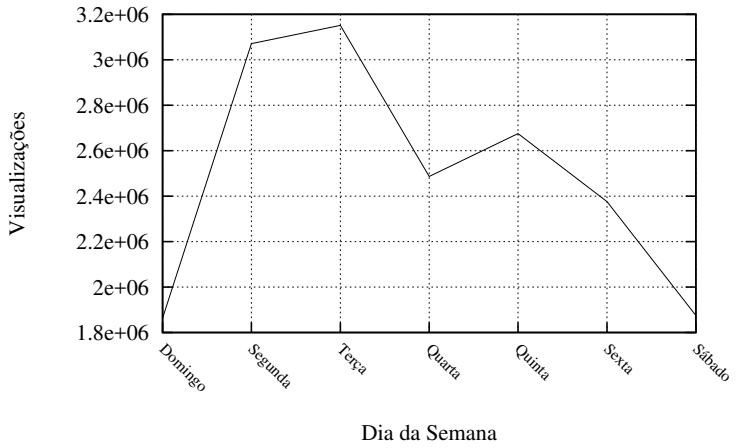


Figura 4.6. Distribuição de visualizações por dia da semana.

As visualizações por dia de semana (Figura 4.6) demonstram, mais uma vez, uma queda de consumo durante os finais de semana (sábado e domingo). Além disso, o início da semana (segunda-feira e terça-feira) possui as maiores quantidades de visualizações agregadas.

Por fim, a Figura 4.7 ilustra uma cCDF da distribuição de usuários distintos que visualizaram um determinado objeto ou mídia. Esse gráfico apresenta valores para até 100 usuários distintos para a melhor visualização da cCDF.

A cCDF traçada na Figura 4.7 ilustra que aproximadamente 50% dos objetos da plataforma Samba Tech são visualizados por até 5 usuários distintos. Além disso, menos de 20% dos vídeos dessa plataforma são visualizados por mais de 50 usuários distintos. Embora não tenha sido demonstrado, a maior quantidade de usuários distintos que acessaram um vídeo no mês de Julho de 2012 chega a um número de aproximadamente 70 mil.

A seguir, apresentamos o modelo e técnica de recomendação propostos para serviços de conteúdo *Web* multimídia.

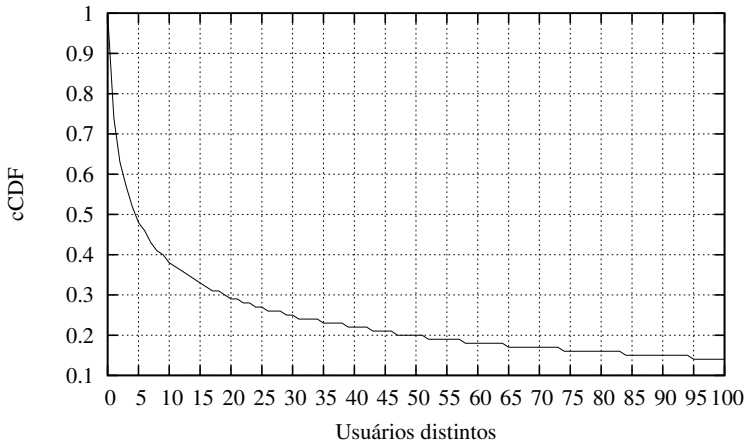


Figura 4.7. cCDF de usuários distintos que visualizaram uma mídia.

Capítulo 5

Recomendação em Serviços de Conteúdo *Web* Multimídia

Apresentamos neste capítulo a técnica proposta para recomendação em serviços de conteúdo *Web* multimídia. A Seção 5.1 descreve um modelo de recomendação para este tipo de serviço. Em seguida, a técnica de recomendação proposta com base nesse modelo é detalhada na Seção 5.2.

5.1 Modelo proposto

Existem diversas maneiras de se modelar o cenário de recomendação para uma plataforma de conteúdo multimídia online, dependendo do objetivo proposto. Uma dessas opções tradicionalmente possui como principal entidade o usuário, e como ele interage com o conteúdo, como demonstrado por Davidson *et al.* [DLL+10].

Propomos neste trabalho uma visão diferente, focada no objeto sendo consumido pelo usuário. Esse objeto corresponde ao conteúdo multimídia em questão, e pode ser um vídeo, imagem, áudio, etc. Tal objeto pode ser consumido de diversas maneiras, em diferentes períodos de tempo ou lugares, por diferentes tipos de usuários. Sendo assim, a modelagem dessas entidades é realizada da seguinte maneira:

- **Objeto**: conteúdo (vídeo, imagem, áudio, etc.) ofertado para o usuário. Possui diversos metadados, como título, descrição e gênero, além de especificações de acordo com o que representa (tempo de duração para vídeos, dimensões para imagens, etc.);
- **Consumo**: situação onde um determinado objeto está sendo consumido. Tal cenário pode ser dividido em:

– **Como/Quando/Onde?**: representa o cenário do consumo, e engloba informações de como (ex.: qual porcentagem de um vídeo foi vista), quando (ex.: a que horas uma imagem foi visualizada) e onde (ex.: de que cidade está sendo realizada o consumo) está sendo gerado o consumo;

– **Quem?**: representa o usuário que está consumindo, assim como toda a gama de informações sobre ele, como sexo, idade, interesses, etc.

Este modelo de consumo de objetos multimídia é representado na Figura 5.1.

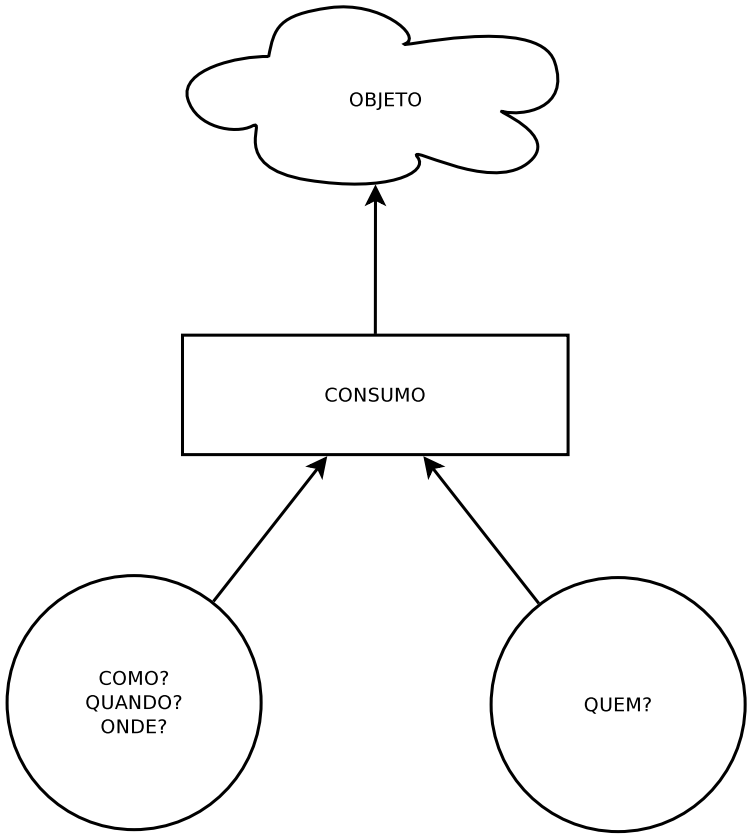


Figura 5.1. Modelo de consumo de objetos multimídia

Esse modelo permite ter uma visão do cenário de consumo de um objeto, com sua divisão em três grande entidades: (i) Objeto, (ii) Como/Quando/Onde é realizado o consumo, e (iii) Quem consome o objeto. Suas combinações permitem diferentes focos de análise. A seguir, apresentamos quatro possibilidades, assim como exemplos de como cada um pode agregar valor ao estudo.

- **Foco 1:** isola-se o objeto (sem nenhuma forma de consumo), permitindo realizar uma série de análises através apenas dos seus metadados (título, descrição, *tags*, etc.) e especificações (duração, dimensão, etc.).

  - Exemplo: quais objetos possuem o mesmo gênero? Quais as *tags* mais utilizadas por um determinado conjunto de mídias? Quais os objetos que mais se assemelham a um objeto dado seus metadados?

- **Foco 2:** isola-se o objeto sendo consumido de uma determinada forma. Essa análise permite avaliar como, quando e de onde um objeto está sendo acessado.

  - Exemplo: Em que horário ocorrem mais acessos a um determinado objeto? Qual a distribuição geográfica de consumo para um determinado conjunto de mídias?

- **Foco 3:** isola-se o objeto sendo consumido por um determinado usuário. Isso permite analisar os diferentes perfis de usuários que consomem um determinado objeto.

  - Exemplo: Dado um objeto, qual a faixa etária de pessoas que o acessam? Um objeto atrai usuários com que tipo de interesses?

- **Foco 4:** por fim, unem-se todas as análises anteriores, formando o fluxo completo da forma com que um determinado objeto é consumido por um usuário.

  - Exemplo: Para uma determinada região, durante as noites dos finais de semana, um objeto atrai quais tipos de usuário?

Com esta segmentação, podemos propor diversas análises, dependendo do foco com que estivermos trabalhando. Uma destas possibilidades é voltada para área de Sistemas de Recomendação [RRSK11, JZFF11]. Em nosso cenário, podemos aplicar esse estudo para a recomendação de conteúdo multimídia direcionada ao usuário, em conjunto com o modelo proposto. Ao avaliar o Foco 1, pode-se recomendar os objetos que mais se assemelham a um determinado objeto, de acordo com seu grupo de metadados e especificações. Por outro lado, com base no Foco 2, pode-se recomendar um conjunto de objetos que, além de considerar esse grupo de metadados e especificações, também avalia o horário e local de consumo (ex.: objetos que mais se assemelham entre si, e que são acessados do mesmo local durante o mesmo período).

A seguir, apresentamos uma técnica de recomendação que possui como base o modelo proposto nesta seção.

5.2 Técnica de Recomendação

Nesta seção, apresentamos uma técnica de recomendação desenvolvida com base no modelo proposto na Seção 5.1, que tem como foco o objeto sendo consumido.

Dentre os problemas mais importantes da área de sistemas de recomendação, citam-se dois: aqueles que estão associados à recomendação de *Melhor Item* e de *Top-N itens* [RRSK11, SK09]. O primeiro consiste em encontrar, para um usuário específico, um item que lhe desperte o maior interesse, comumente definido a partir das classificações (*ratings*) realizadas nos itens da base de dados. Quando tais classificações não estão disponíveis, e apenas a lista de compras ou acessos de cada usuários é conhecida, o problema de se encontrar o item mais interessante se transforma na tarefa de recomendar a um usuário uma lista de itens contendo N objetos que possam interessá-lo.

Em nosso cenário de serviços de conteúdo multimídia, consideramos que a classificação de objetos é um aspecto mais raro e difícil de se obter. Ao avaliarmos serviços como Youtube, Flickr ou até mesmo a plataforma da Samba Tech, a visualização ou não de um determinado objeto é a informação mais facilmente obtida e confiável. Dessa forma, a ideia principal de nossa técnica se baseia na recomendação de uma lista de potenciais objetos, o que aborda métodos baseados em *Top-N* itens.

Seguindo a metodologia da Seção 5.1, apresentamos uma técnica de recomendação com foco no objeto. Nossa proposta é a geração de uma lista de potenciais itens com base na similaridade entre objetos. Considerando um determinado item de nossa base de dados, realizamos a sua comparação com todos os itens restantes dessa mesma base. Feito isso, nossa técnica recomenda uma lista dos N itens que mais se assemelham a esse determinado item.

Um dos principais aspectos de nossa técnica é o método utilizado para a comparação entre objetos. Para isso, utilizamos uma combinação de dimensões que, seguindo novamente a metodologia da Seção 5.1, podem envolver o objeto e/ou seu consumo. Tais dimensões podem ser definidas a partir do:

- ***Objeto***: agrupa apenas metadados e especificações do objeto. Exemplos de dimensões: título, descrição e gênero (metadados), tempo de duração e dimensões (especificações).

- ***Como/Quando/Onde*** *o objeto é consumido*: agrupa informações relativas ao consumo. Exemplos de dimensões: popularidade (quantidade de visualizações), localização e horário do consumo.

- ***Quem*** *está consumindo o objeto*: agrupa informações do usuário. Exemplos de dimensões: sexo, idade, interesses.

Dessa maneira, a similaridade entre objetos ocorre utilizando-se um subconjunto dessas dimensões para fins de comparação. Formalmente, nossa técnica pode ser descrita da seguinte forma: considerando um conjunto O de objetos, geramos sua lista L_o de similaridades:

$$\forall o \in O \rightarrow L_o = \{\forall x \in \ O \rightarrow sim(o, x)\} \tag{5.1}$$

A função de similaridade $sim(o_1, o_2)$ é calculada a partir das dimensões citadas. Suponhamos que cada objeto o possua um conjunto D_o contendo m dimensões, sejam elas do objeto e/ou do seu consumo. Cabe a essa função realizar a comparação entre D_{o_1} e D_{o_2}. Esse processo pode ser realizado considerando que cada conjunto D de dimensões é um vetor. Dessa forma, a função de similaridade $sim(o_1, o_2)$ retorna o produto escalar entre esses dois vetores:

$$o \rightarrow \vec{o} = \vec{D_o} = \{D_1, D_2, \ldots, D_m\} \tag{5.2}$$

$$sim(o_1, o_2) \rightarrow \vec{o_1} \cdot \vec{o_2} \tag{5.3}$$

Em seguida, essa lista de similaridades é ordenada, e uma lista R_o (*rankings*) é gerada contendo os N itens mais similares a cada um dos objetos o:

$$\forall o \in O \rightarrow LS_o = sort(L_o) \tag{5.4}$$

$$\forall o \in O \rightarrow R_o = \{LS_1, LS_2, ..., LS_N\} \tag{5.5}$$

É importante ressaltar que, devido ao foco de nossa técnica de recomendação, a geração dos potenciais itens recomendados é realizada para cada objeto. Sendo assim, tal recomendação somente pode ser realizada a partir do acesso a um determinado objeto. Em outras palavras, só é possível recomendarmos objetos após o acesso a um primeiro item. Com isso, não é possível solucionar o problema de *cold start*, que se refere à dificuldade de gerar recomendações de novos itens ou para novos usuários, uma vez que existe pouca informação sobre eles [SPUP02].

No próximo capítulo apresentamos como essa técnica de recomendação é aplicada em nosso cenário.

Capítulo 6

Técnica de Recomendação, Validação e Análise dos Resultados

Neste capítulo apresentamos a técnica de recomendação proposta no capítulo 5, aplicada no ambiente da plataforma Samba Tech. Os detalhes da aplicação são apresentados na Seção 6.1. A função de similaridade entre itens utilizada nessa aplicação é explicada na Seção 6.2. Em seguida, a Seção 6.3 apresenta o método utilizado como *baseline* para a comparação de nossa aplicação. A explicação da validação de nossa aplicação se encontra na Seção 6.4, seguida pelos resultados de nossa técnica de recomendação aplicada na Seção 6.5.

6.1 Aplicação da Técnica de Recomendação

Nesta seção apresentamos a aplicação da técnica apresentada na Seção 5.2, em conjunto com o modelo exposto na Seção 5.1, com o intuito de recomendar conteúdo multimídia em nosso cenário. Essa aplicação será realizada sobre os dados coletados da plataforma Samba Tech. Devido à sua característica (Tabela 4.1), nossa aplicação envolverá apenas vídeos como objetos, uma vez que esses representam quase a totalidade de seus dados.

Para isso, inicialmente é necessário definir como será realizada a similaridade entre objetos (ou vídeos), ou seja, como implementaremos a função $sim(o_1, o_2)$. O modelo proposto para o cenário de recomendação com foco no objeto oferece diferentes focos de análise, de acordo com o objetivo do estudo. Dessa forma, abordaremos nesse trabalho a aplicação de cada um desses focos em nossa técnica de recomendação.

Essa abordagem será realizada com base nas três entidades citadas no modelo de recomendação: o objeto, como/quando/onde esse objeto é consumido, e quem realiza

tal consumo. Cada uma dessas entidades possui um conjunto de dimensões específicos, que serão utilizadas para a nossa função de similaridade e geração dos *rankings*.

Entretanto, os dados coletados pela ferramenta STTM não contêm informações sobre a entidade usuário. Tais informações são, de fato, de difícil obtenção, uma vez que envolvem o perfil do usuário e toda gama de informações a seu respeito. Portanto, não utilizamos essa entidade para a aplicação da técnica de recomendação proposta. As dimensões utilizadas para cada entidade com base na plataforma Samba Tech são:

- **Objeto**:
    1. Gênero: corresponde ao gênero do objeto. Os possíveis valores são: Animais, Ciência, Comédia, Entretenimento, Esportes, Filmes, Música, Pessoas e Política, além do caso onde nenhum valor é atribuído ao objeto, que resulta na opção "Sem Gênero".
    2. Tempo de duração: representa o tempo de duração (em segundos) do vídeo em questão.
    3. Projeto: representa o projeto em que o vídeo se encontra. Cada objeto da plataforma Samba Tech pertence a um projeto, que por sua vez, pertence a um determinado cliente. O projeto de um cliente corresponde a um programa da emissora. Um exemplo de projeto seria o *Jornal da SBT* que pertence ao cliente SBT.
    4. *Tags*: lista de etiquetas ou rótulos (*strings*) inseridas para classificação de cada objeto.
- **Como/Quando/Onde**:
    1. Popularidade: quantidade de visualização de um determinado vídeo.
    2. Horário de acesso: momento em que o vídeo foi acessado.
    3. Dispositivo de acesso (*Device*): tipo do dispositivo onde o objeto foi acessado. Os possíveis valores informam apenas se o dispositivo é móvel (e.g., celular e *tablets*) ou não (e.g., computadores pessoais e *notebooks*).
    4. Origem do acesso (*Referrer*): URL (*Uniform Resource Locator*) do endereço que originou o acesso à página onde o *player* se encontra.
    5. Localização de acesso: região de origem do acesso.

A combinação dessas dimensões para o método de similaridade entre objetos será realizada com base nos diferentes focos apresentados na metodologia da Seção 5.1.

A partir das informações de duas entidades, é possível ter diferentes abordagens que geram diferentes recomendações. Especificamente, existem as seguintes:

- **Recomendação 1 (REC1)**: baseada no *Foco 1*, utiliza apenas as dimensões do objeto para a função de similaridade;
- **Recomendação 2 (REC2)**: baseada no *Foco 2*, utiliza apenas as dimensões do consumo do objeto para a função de similaridade;
- **Recomendação 3 (REC3)**: baseada no *Foco 4*, utiliza a combinação das dimensões do objeto e de seu consumo para a função de similaridade. Ressaltamos que não incluímos informações do usuário, já que não dispomos de tais dados.

O foco deste trabalho está nessas três possibilidades de recomendação. O intuito dessa abordagem é demonstrar que, agregando-se mais informações sobre o consumo de um objeto, é possível gerar uma lista de itens similares de melhor qualidade. Definimos, a partir disso, a seguinte hipótese:

> *Quanto mais dimensões do consumo do item utilizarmos para a comparação entre itens, mais refinada será a geração de nossos* rankings, *e consequentemente, faremos uma recomendação melhor.*

6.2 Similaridade entre itens

Após a definição das dimensões utilizadas no processo de recomendação, é necessário escolher o método para a combinação dessas dimensões. Tal conjunto pode ser tratado como um vetor, e dessa forma, existem diversas técnicas para a sua comparação [TSK05]. Para esse propósito, é utilizada a Similaridade de Cosseno [SKKR01], uma das técnicas mais conhecidas na área de Mineração de Dados. Nesse caso, a similaridade de dois itens é medida a partir do ângulo formado pelos dois vetores em questão em um espaço m dimensional. Formalmente, em uma matriz de $m \times n$ dimensões, a similaridade entre o_1 e o_2 definida por $sim(o_1, o_2)$ é calculada de acordo com a Equação 6.1:

$$sim(o_1, o_2) = cos(\vec{o_1}, \vec{o_2}) = \frac{\vec{o_1}.\vec{o_2}}{\|o_1\| \times \|o_2\|} \tag{6.1}$$

A Similaridade de Cosseno é utilizada como principal método de comparação entre objetos e suas dimensões. Porém, cada uma dessas dimensões avaliadas possuem suas particularidades, e exigem um processo único para a sua combinação. Um exemplo

dessa necessidade é a combinação da dimensão referente à localização de acesso, que possui um conjunto diversificado de possibilidades, o que dificulta a sua comparação.

O cálculo individual de cada dimensão é demonstrado a seguir. Tais métodos utilizam a combinação de vários conceitos, e o valor calculado para a similaridade entre cada dimensão está no intervalo $[0, 1]$. Esse processo utiliza a própria similaridade de cosseno para a comparação de vetores de inteiros, além do Coeficiente de Similaridade de Jaccard [TSK05], que calcula a diversidade entre conjuntos. Normalmente utilizada para a comparação de conjuntos de textos (*strings*), recebe como entrada duas amostras A e B, e pode ser definido de acordo com a Equação 6.2:

$$JACCARD(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.2}$$

A similaridade entre cada uma das dimensões estudadas é definida a seguir.

1. Gênero

   - Tipo: texto (*string*)
   - Similaridade entre G_{o_1} e G_{o_2}:

$$G_{o_1} \times G_{o_2} = \begin{cases} 1,0, & \text{se } G_{o_1} = G_{o_2} \text{ e } G_{o_1} \neq \text{Sem Gênero e } G_{o_2} \neq \text{Sem Gênero} \\ 0,0, & \text{se } G_{o_1} \neq G_{o_2} \text{ e } G_{o_1} \neq \text{Sem Gênero e } G_{o_2} \neq \text{Sem Gênero} \\ 0,5, & G_{o_1} = \text{Sem Gênero ou } G_{o_2} = \text{Sem Gênero} \end{cases}$$

2. Tempo de duração

   - Tipo: inteiro (quantidade de segundos)
   - Similaridade entre T_{o_1} e T_{o_2}:

$$T_{o_1} \times T_{o_2} = \begin{cases} \frac{T_{o_1}}{T_{o_2}}, & \text{se } T_{o_2} > T_{o_1} \\ \frac{T_{o_2}}{T_{o_1}}, & \text{caso contrário} \end{cases}$$

3. Projeto

   - Tipo: um inteiro identificando o Projeto (p_o), e um inteiro identificando o Cliente (c_o)

- Similaridade entre P_{o_1} e P_{o_2}:

$$P_{o_1} \times P_{o_2} = \begin{cases} 1,0, & \text{se } p_{o_1} = p_{o_2} \text{ e } c_{o_1} = c_{o_2} \\ 0,5, & \text{se } c_{o_1} = c_{o_2} \\ 0,0, & \text{caso contrário} \end{cases}$$

4. *Tags*

   - Tipo: lista de textos (*array* de *string*)
   - Similaridade entre TG_{o_1} e TG_{o_2}:

$$TG_{o_1} \times TG_{o_2} = JACCARD(\vec{TG}_{o_1}, \vec{TG}_{o_2})$$

5. Popularidade

   - Tipo: inteiro (quantidade de visualizações)
   - Similaridade entre V_{o_1} e V_{o_2}:

$$V_{o_1} \times V_{o_2} = \begin{cases} \frac{V_{o_1}}{V_{o_2}}, & \text{se } V_{o_2} > V_{o_1} \\ \frac{V_{o_2}}{V_{o_1}}, & \text{caso contrário} \end{cases}$$

6. Horário de acesso

   - Tipo: lista de inteiros
   - Descrição: O valor da posição i corresponde à quantidade de visualizações na hora i, sendo que $i \in [0, 24)$
   - Similaridade entre H_{o_1} e H_{o_2}:

$$H_{o_1} \times H_{o_2} = cos(\vec{H_{o_1}}, \vec{H_{o_2}})$$

7. Dispositivo de acesso (*Device*)

   - Tipo: lista de inteiros
   - Descrição: lista contém sempre 2 elementos. A primeira posição contém a quantidade de visualizações realizada a partir de dispositivos móveis. A segunda posição corresponde às visualizações de dispositivos que não são móveis.

- Similaridade entre D_{o_1} e D_{o_2}:

$$D_{o_1} \times D_{o_2} = cos(\vec{D_{o_1}}, \vec{D_{o_2}})$$

8. Origem do acesso (*Referrer*)

   - Tipo: lista de inteiros
   - Descrição: Realizamos a contabilização dos 10 domínios[1] que mais possuem visualizações, e geramos uma lista com esses valores ordenados.
   - Similaridade entre R_{o_1} e R_{o_2}:

$$R_{o_1} \times R_{o_2} = cos(\vec{R_{o_1}}, \vec{R_{o_2}})$$

9. Localização de acesso

   - Tipo: lista de inteiros
   - Descrição: Realizamos a contabilização das 10 regiões que mais possuem visualizações, e geramos uma lista com esses valores ordenados.
   - Similaridade entre L_{o_1} e L_{o_2}:

$$L_{o_1} \times L_{o_2} = cos(\vec{L_{o_1}}, \vec{L_{o_2}})$$

6.3 Algoritmo *Baseline*

Em adição às três propostas de recomendação detalhadas na Seção 6.1, nosso estudo inclui a comparação dessas técnicas com um modelo base, conhecido como *baseline*. Esse processo é relevante para avaliarmos a eficácia de nossa técnica de recomendação proposta.

Para isso, nós aplicamos um algoritmo conhecido para a geração de uma lista de potenciais itens recomendados. A técnica utilizada foi o WRMF, ou *Weighted Regularized Matrix Factorization* [HKV08, PZC+08], considerada na literatura como uma técnica estado-da-arte de fatorização de matrizes para a área de *Collaborative Filtering* em *Top-N* itens [DADSTN12].

[1] Extraído a partir da URL, o domínio é um nome que serve para localizar e identificar conjuntos de computadores na Internet. O nome de domínio foi concebido com o objetivo de facilitar a memorização dos endereços de computadores na Internet. Sem ele, teríamos que memorizar uma sequência grande de números.

A geração de *rankings* a partir do WRMF considera apenas o conjunto de visualizações dos usuários para a criação da matriz de fatorização. Sua execução utilizou a biblioteca chamada *MyMediaLite* [GRFST11], que oferece a implementação de diversas técnicas de sistemas de recomendação.

Uma peculiaridade dessa ferramenta é a geração dos seus *rankings*. Nossa proposta de recomendação propõe o foco no objeto, com a recomendação de itens mais similares a um determinado objeto. Em outras palavras, a recomendação de itens é realizada para um determinado objeto (Seção 5.2). O método WRMF implementado pela biblioteca *MyMediaLite* também resulta em uma lista de itens recomendados, mas tal recomendação é realizada para um usuário, e não um objeto. Essa solução não impacta em nossa comparação, como detalhado nas seções seguintes.

6.4 Validação da Técnica de Recomendação

O método de validação utilizado para a avaliação de técnicas de recomendação é de crucial importância para o estudo, sendo propostos inúmeros métodos para esse propósito [SG11]. O nosso cenário considera o objeto em foco para a geração da lista de itens recomendados. Porém, apenas com a base de sessões fornecida pela ferramenta STTM, não é possível realizar a recomendação de fato ao usuário. Sendo assim, é necessário separar nossa base de dados em uma base de treino e uma base de testes. O primeiro, corresponde à base utilizada para a geração dos *rankings* de recomendação. A segunda, corresponde à base que será utilizada para a validação da recomendação gerada.

A partir da escolha das bases de treino e de testes, é necessário definir o método utilizado para a validação. O método desenvolvido para esse propósito é descrito a seguir.

6.4.1 Método de Validação

O método de validação proposto possui como entrada os *rankings* gerados a partir da base de treino, e as visualizações dos usuários da base de teste. Nosso desafio de validação é relacionado ao problema de *cold start* citado na Seção 5.2: é necessário que o usuário visualize pelo menos um objeto para gerar a recomendação de itens a partir dos *rankings* do objeto visualizado.

O método proposto se baseia na divisão da base de testes em duas: (i) a primeira é utilizada para a escolha dos *rankings* dos itens visualizados, assim como a seleção de N itens recomendados com base nos N itens de maiores *rankings* (ou seja, que são

mais similares ao objeto visualizado); (ii) a segunda corresponde aos itens visualizados pelo usuário que serão testados para a eficácia da recomendação.

Considerando que um usuário u possua um conjunto V contendo X objetos visualizados:

$$u \rightarrow V = \{v_1, v_2, \ldots, v_X\}$$

O conjunto V é então divido em dois conjuntos VR e VT. O primeiro conjunto contém os objetos utilizados para escolha dos *rankings*. O segundo conjunto representa os objetos que serão testados para avaliar a eficácia da recomendação.

Porém, ainda é necessário escolher o ponto de divisão da base testes, assim como os tamanhos dos conjuntos VR e VT. Para isso, são propostas quatro abordagens que trabalham com uma divisão dessa base. Supondo que o ponto de divisão ocorra no i-ésimo objeto visualizado, são definidas quatro opções:

$$\text{Passado Recente-Futuro Recente} \rightarrow \{v_1, v_2, \ldots, \underbrace{v_i}_{VR}, \underbrace{v_{i+i}}_{VT}, v_{i+2}, \ldots, v_X\}$$

$$\text{Passado Recente-Futuro Completo} \rightarrow \{v_1, v_2, \ldots, \underbrace{v_i}_{VR}, \underbrace{v_{i+i}, v_{i+2}, \ldots, v_X}_{VT}\}$$

$$\text{Passado Completo-Futuro Recente} \rightarrow \{\underbrace{v_1, v_2, \ldots, v_i}_{VR}, \underbrace{v_{i+i}}_{VT}, v_{i+2}, \ldots, v_X\}$$

$$\text{Passado Completo-Futuro Completo} \rightarrow \{\underbrace{v_1, v_2, \ldots, v_i}_{VR}, \underbrace{v_{i+i}, v_{i+2}, \ldots, v_X}_{VT}\}$$

Dessa forma, o ponto de divisão i é realizado de forma a percorrer todas as visualizações do usuário u. Ou seja, ocorrem variações de forma que o tamanho dos conjuntos VR e VT variem de 1 a $X - 1$.

Por fim, é necessário definir como é feita a recomendação de itens a partir da escolha dos *rankings* gerados dos objetos do conjunto VR. Essa geração é realizada a partir da combinação dos *rankings* de todos os objetos contidos em VR. Caso um objeto ocorra mais de uma vez, utilizamos o maior valor de sua similaridade. O resultado dessa escolha é uma lista IR de tamanho N, contendo os itens mais similares ao conjunto de objetos formado por VR.

A avaliação da eficácia de nossa técnica a partir dos conjuntos IR e VT é desenvolvida a partir de inúmeras métricas conhecidas na literatura. A seguir, detalhamos cada uma dessas métricas.

6.4.2 Métricas para Validação

Com a definição do método de validação na Seção 6.4.1, devemos avaliar a eficácia de nossa recomendação. Para esse propósito, utilizamos uma série de métricas descritas a seguir.

6.4.2.1 Precisão/Revocação

Em um cenário onde os objetos não possuem uma classificação, sendo que apenas a lista de itens acessados por um usuário está disponível, são encontradas diversas dificuldades na medição da acurária da predição de classificação. Dessa forma, nosso problema se resume a medir a eficácia da recomendação de uma lista de potenciais itens para o usuário. Dessa maneira, não utilizamos em nossa pesquisa duas das métricas mais conhecidas na área de sistemas de recomendação: MAE (*Mean Absolute Error*) e RMSE (*Root Mean Squared Error*). Em nosso contexto, as métricas mais aconselhadas para a medição da acurária de nossa técnica de recomendação são *Precisão*[2] e *Revocação*[2] [DK11].

Formalmente, essas métricas são definidas da seguinte maneira: considerando que o usuário u possua um conjunto de testes $T(u)$ contendo os objetos que ele realmente acessou, e dado um conjunto $R(u)$ contendo os objetos recomendados para o usuário em questão, as Equações 6.3 e 6.4 definem Precisão e Revocação respectivamente:

$$\text{Precisão} = \frac{T(u) \cap R(u)}{R(u)} \tag{6.3}$$

$$\text{Revocação} = \frac{T(u) \cap R(u)}{T(u)} \tag{6.4}$$

Aplicando essa definição em nosso estudo, de acordo com o método de validação definido na Seção 6.4.1, temos que $T(u) = VT$ e $R(u) = IR$. Porém, nossa validação aplicada utiliza uma variação da métrica de Precisão: caso o usuário tenha visto pelo menos um vídeo do nosso conjunto de itens recomendados IR, a precisão de nossa recomendação foi 100%, caso contrário, nossa recomendação obteve 0% de precisão.

Essa variação é utilizada por diversos motivos: como a base de dados possui a visualização de todos os vídeos da plataforma Samba Tech, nossa recomendação de vídeos não é apresentada de fato para os usuários. Dessa forma, precisamos inferir se o usuário visualizou ou não um dos vídeos que teríamos recomendado utilizando a técnica proposta. Isso é feito verificando se o usuário assistiu ou não a um dos vídeos do conjunto IR. Porém, a não visualização de um vídeo recomendado não representa

[2]Mais comumente utilizadas em seus termos em inglês: *Precision* e *Recall*

que a técnica foi falha, pois os vídeos não foram de fato apresentados como opção ao usuário. O valor "zero" escolhido para a métrica nesse caso foi utilizado com o intuito de comparação dos nossos resultados. Da mesma forma, caso um usuário tenha visualizado pelo menos um vídeo pertencente a IR, tem-se uma precisão de 100%. Isso é feito, pois, em nosso cenário de vídeos online, podemos considerar que nossa recomendação teve sucesso pelo fato do usuário ter acessado um dos N objetos recomendados, não havendo a necessidade do usuário acessar os N itens para que tenhamos 100% de precisão. Um exemplo para um melhor entendimento dessa abordagem é o Youtube: ao finalizar uma visualização, são recomendados um conjunto de novos vídeos. Caso o usuário acesse pelo menos um desses itens, pode-se dizer que a recomendação obteve sucesso.

Dessa forma, a métrica de Precisão em nosso cenário é redefinida como:

$$\text{Precisão} = \begin{cases} 100\%, & \text{se } |VT \cap IR| \geq 1 \\ 0\%, & \text{caso contrário} \end{cases} \tag{6.5}$$

6.4.2.2 Precisão*/Revocação*

Complementar às métricas de Precisão e Revocação citadas na Seção 6.4.2.1, uma segunda variação proposta no trabalho de Cremonesi *et al.* [CKT10] é utilizada. Sua metodologia realiza uma extensão das métricas de Precisão e Revocação como descrito a seguir:

1. Considere que, na base de testes, o usuário u tenha classificado o item i;
2. Aleatoriamente são escolhidos 1000 itens adicionais que não foram ainda classificados pelo usuário u. Assumimos que esses itens não são de interesse de u;
3. Prevemos a classificação para o item i e esses 1000 itens adicionais;
4. Ordena-se a lista contendo esses 1001 itens. Seja p a posição do item i após a ordenação desses itens. O melhor resultado possível ocorre quando o item i tem melhor classificação do que todos os outros itens adicionais (i.e., $p = 1$);
5. Realiza-se uma recomendação do tipo *Top-N* nessa lista de 1001 itens. Se $p \leq N$, é considerado um acerto (*hit*), caso contrário, uma falha (*miss*).

Aplicando-se essa metodologia em nosso cenário, o seguinte processo é definido:

1. Considere, para um determinado usuário u, o conjunto de itens recomendados IR a partir de um único item i, e seus itens visualizados VT (Seção 6.4.1) de nossa base de testes;

2. Aleatoriamente são escolhidos a vídeos adicionais que, conhecidamente, não foram ainda visualizados pelo usuário u;

3. Prevemos os *rankings* para o conjunto IR e para esse conjunto adicional contendo a vídeos;

4. Ordena-se a lista contendo todo esses conjuntos de itens;

5. Realiza-se uma recomendação de N itens a partir dessa lista ordenada;

6. Calcula-se a precisão e revocação como demonstrados na Seção 6.4.2.1.

Após diversos testes realizados, descobriu-se que o melhor valor para o conjunto de itens aleatórios é $a = 500$. Apesar de ser um valor diferente do utilizado na metodologia original, o valor escolhido cumpre satisfatoriamente o propósito da metodologia original para o nosso problema e cenário de investigação e validação.

6.4.2.3 *Rank Score*

Proposto no trabalho de Shani e Gunawardana [SG11], a métrica de *Rank Score* estende a métrica de revocação para considerar a posição do item recomendado no *ranking*. Possui grande importância em sistemas de recomendação, uma vez que itens com *ranking* baixo podem ser negligenciados pelos usuários.

Para o cálculo da métrica de *Rank Score*, é necessário considerar as seguintes premissas:

- Considere como h o conjunto de vídeos corretamente recomendados;
- Seja a função $rank(i)$ definida para o cálculo da posição do item i no *ranking*;
- Seja T o conjunto de itens recomendados;
- Seja α definido no trabalho como *ranking half life* (i.e., um fator de redução exponencial).

Dessa forma, o *Rank Score* é calculado de acordo com as Equações 6.6, 6.7 e 6.8:

$$\text{Rank Score} = \frac{rankscore_p}{rankscore_{max}} \tag{6.6}$$

$$rankscore_p = \sum_{i \in h} 2^{-\frac{rank(i)-1}{\alpha}} \tag{6.7}$$

$$rankscore_{max} = \sum_{i=1}^{|T|} 2^{-\frac{i-1}{\alpha}} \tag{6.8}$$

Para os testes realizados nessa dissertação, foi-se utilizado o valor 2 para a constante α, assim como considerado em exemplos do trabalho de Shani e Gunawardana.

6.4.2.4 *Normalized Discounted Cumulative Gain*

Assim como a métrica de *Rank Score* (Seção 6.4.2.3), é possível utilizar uma outra métrica que considera a posição do item no *ranking*: o nDCG, ou *Normalized Discounted Cumulative Gain* [JK02]. Utilizando uma escala de relevância gradual de documentos em um conjunto de resultados de uma máquina de busca, o nDCG mede a utilidade ou ganho de um documento baseado na sua posição no *ranking*.

O seu cálculo considera as seguintes variáveis:

- pos a posição de relevância acumulada;
- h o conjunto de itens corretamente recomendados;
- rel_i retorna a relevância da recomendação na posição i.

A fórmula para o cálculo do nDCG é definida pelas Equações 6.9, 6.10 e 6.11:

$$nDCG_{pos} = \frac{DCG_{pos}}{IDCG_{pos}} \tag{6.9}$$

$$DCG_{pos} = rel_1 + \sum_{i=2}^{pos} \frac{rel_i}{\log_2 i} \tag{6.10}$$

$$IDCG_{pos} = rel_1 + \sum_{i=2}^{|h|-1} \frac{rel_i}{\log_2 i} \tag{6.11}$$

Utilizamos como base a implementação utilizada pela biblioteca *MyMediaLite*[3] [GRFST11].

6.5 Resultados

Nesta seção apresentamos os resultados da aplicação da técnica de recomendação proposta nessa dissertação. Inicialmente, a Seção 6.5.1 descreve os dados utilizados nesses

[3]Mais detalhes da implementação podem ser encontradas em: `www.github.com/zenogantner/MyMediaLite/blob/master/src/MyMediaLite/Eval/Measures/NDCG.cs`

resultados. Em seguida, os resultados segmentados com base nas métricas propostas são apresentados na Seção 6.4.2.

6.5.1 Descrição dos resultados

Com uma vasta base de dados, a primeira tarefa para a execução dos experimentos é o particionamento dos dados para treinamento e testes. Dessa forma, a base é separada conforme a Tabela 6.1.

	Período utilizado
Base de treino	01/07/2012 a 07/07/2012
Base de testes	08/07/2012 a 14/07/2012

Tabela 6.1. Base de dados utilizada nos experimentos.

Em suma, foi escolhida a primeira semana de nossa base para treino, ou seja, para a geração dos *rankings*, e a semana seguinte para a base de testes, onde serão aplicadas as validações descritas na Seção 6.4.1. Ressaltamos que serão mostrados resultados de apenas duas semanas de nossa base de dados, uma vez que tal amostra representa boa parcela de nossa base de dados, contemplando grande volume de dados (milhões de sessões de usuário), o que consiste de um método consistente de validação experimental para sistemas de recomendação.

Alguns aspectos importantes devem ser considerados para o melhor entendimento das análises a seguir. Desse modo, para cada métrica avaliada, apresentamos os resultados para as quatro validações propostas: Passado Recente - Futuro Recente, Passado Recente - Futuro Completo, Passado Completo - Futuro Recente, Passado Completo - Futuro Completo. Entretanto duas exceções são apresentadas:

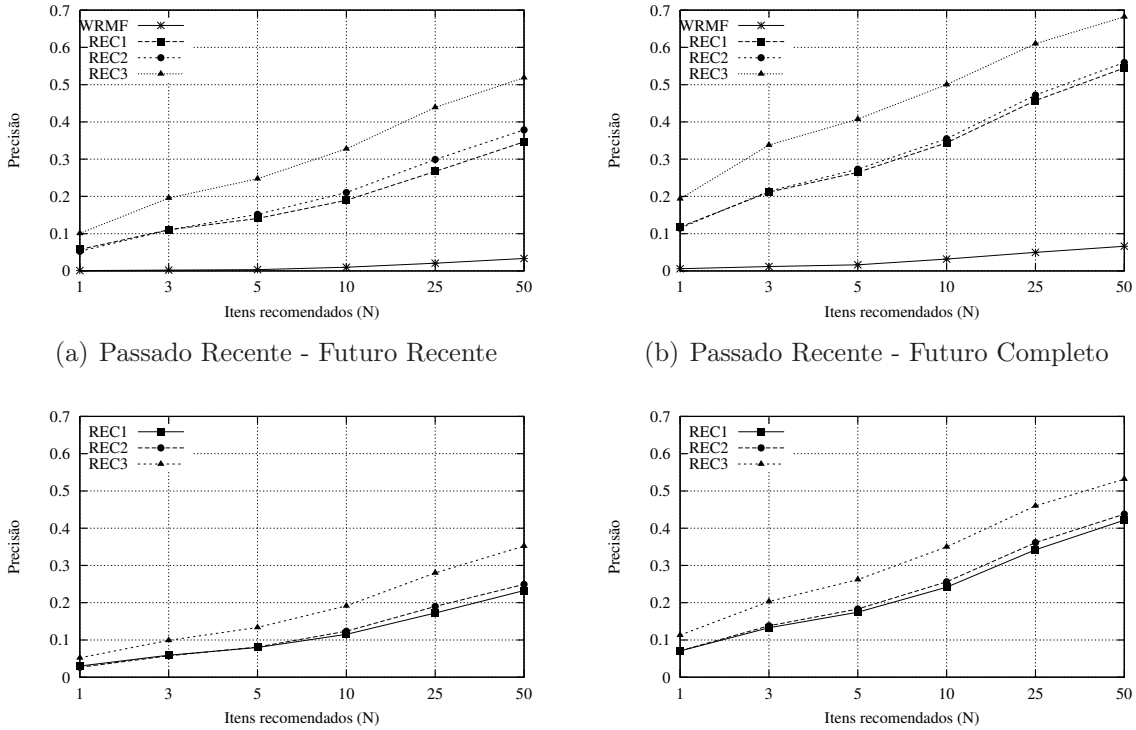
1. O método de *baseline* utilizado possui uma característica diferente das nossas recomendações propostas, uma vez que os *rankings* são recomendados por usuário. Dessa forma, realizamos as validações apenas com base na variação do conjunto VT, considerando o Futuro Recente (apenas primeiro item de VT) e Futuro Completo (todos os itens de VT).

2. Para as métricas de precisão*e revocação*, não apresentamos resultados que envolvam a validação de Passado Completo. Isso ocorre pois, como explicado na Seção 6.4.2.2, é utilizado um conjunto IR de objetos recomendados adicionalmente aos itens aleatórios. Caso utilizássemos uma combinação de *rankings* para a geração de IR (utilizando a validação com Passado Completo), precisaríamos

realizar o mesmo processo para os itens aleatórios, o que dificultaria o processo final de uma maneira geral.

6.5.2 Resultados para precisão

Os resultados relacionados à métrica de precisão, descrita na Seção 6.4.2.1, são expostos na Figura 6.1.



(a) Passado Recente - Futuro Recente

(b) Passado Recente - Futuro Completo

(c) Passado Completo - Futuro Recente

(d) Passado Completo - Futuro Completo

Figura 6.1. Resultados da aplicação da técnica de recomendação: precisão.

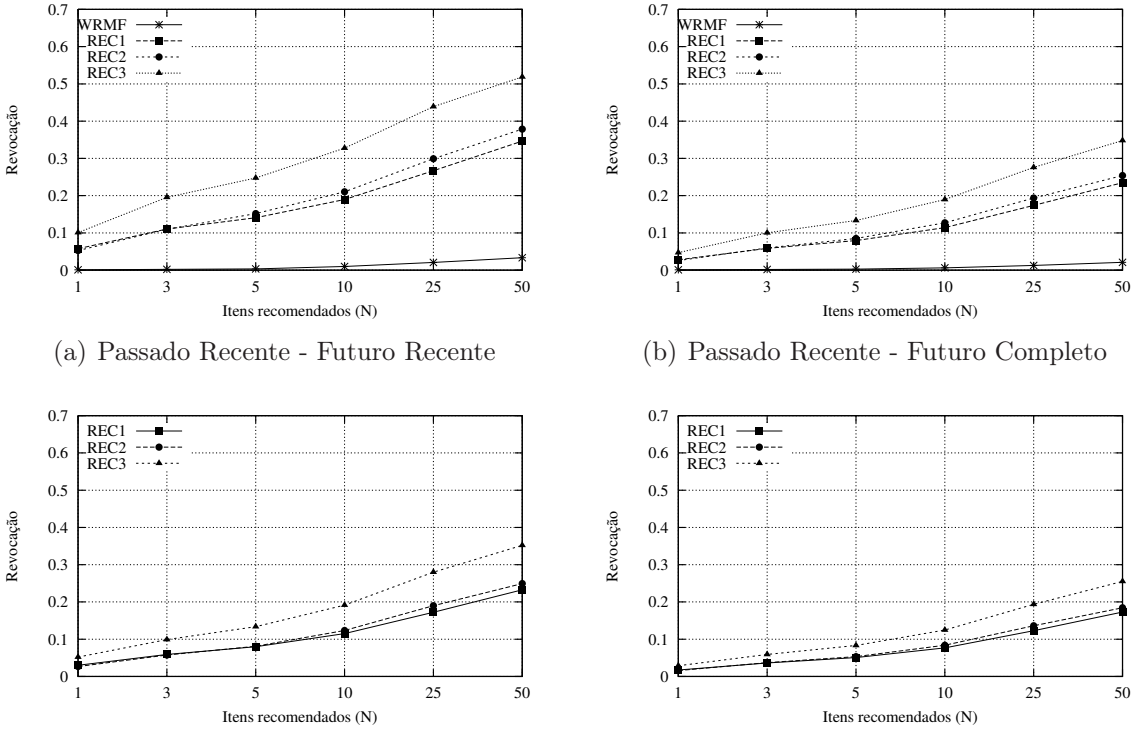
A partir dos gráficos da Figura 6.1, é possível perceber que o valor da precisão aumenta ao recomendar mais itens. De uma maneira geral, a técnica de recomendação utilizada como *baseline* obteve resultados menores em comparação com nossas três recomendações. Entre essas, os melhores valores de precisão foram obtidos ao aplicar a recomendação REC3, que combina informação do objeto e de seu consumo. O melhor valor obtido considera 50 itens recomendados, seguindo a validação de Passado Recente - Futuro Completo (Figura 6.1(b)), chegando a uma precisão de 68,21%.

Com base nas quatro validações apresentadas, os resultados utilizando nossas recomendações foram menores na avaliação de todos os vídeos visualizados pelo usuário

para gerar os *rankings*, e testar apenas com o próximo vídeo acessado (Passado Completo - Futuro Recente, Figura 6.1(c)), enquanto os melhores resultados foram obtidos no cenário oposto (Passado Recente - Futuro Completo, Figura 6.1(b)). Esse comportamento ocorre pois a quantidade de itens utilizados para o cálculo da precisão é maior nesse último caso.

6.5.3 Resultados para revocação

Os resultados para a revocação, métrica detalhada também na Seção 6.4.2.1, podem ser vistos na Figura 6.2.



(a) Passado Recente - Futuro Recente

(b) Passado Recente - Futuro Completo

(c) Passado Completo - Futuro Recente

(d) Passado Completo - Futuro Completo

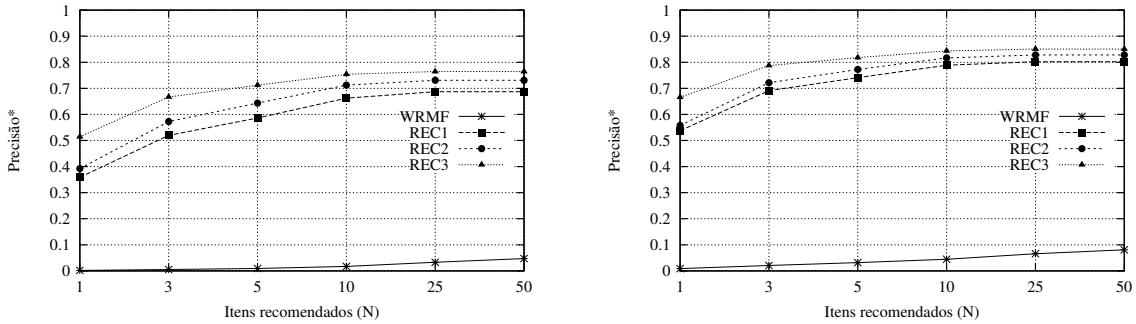
Figura 6.2. Resultados da aplicação da técnica de recomendação: revocação.

Os gráficos da Figura 6.2 demonstram que, dentre os vídeos visualizados pelo usuário, a quantidade de itens recomendados seguindo o nosso modelo aumenta. Porém, esse crescimento é pouco significativo ao avaliar a técnica WRMF, que obteve os menores valores de revocação. Comparando-se as técnicas de recomendação propostas nessa dissertação, percebe-se que as revocações calculadas ao avaliar o objeto e seu consumo separadamente (REC1 e REC2, respectivamente) são similares.

As quatro validações distintas utilizadas demonstram valores bem próximos, excluindo o caso do Passado Recente - Futuro Recente (Figura 6.2(a)). Nesse gráfico, os valores de nossa recomendação são os maiores, chegando a uma revocação de 51,87%. Esse comportamento pode ser explicado pelo fato de, nesse cenário, o conjunto de itens visualizados pelo usuário ser o menor possível (apresentando apenas um vídeo) e a quantidade de itens recomendados ser crescente, o que impacta diretamente no cálculo da métrica em questão.

6.5.4 Resultados para precisão*

A Figura 6.3 ilustra os resultados para a variação da métrica de precisão utilizada, nomeada de precisão*(Seção 6.4.2.2).



(a) Passado Recente - Futuro Recente (b) Passado Recente - Futuro Completo

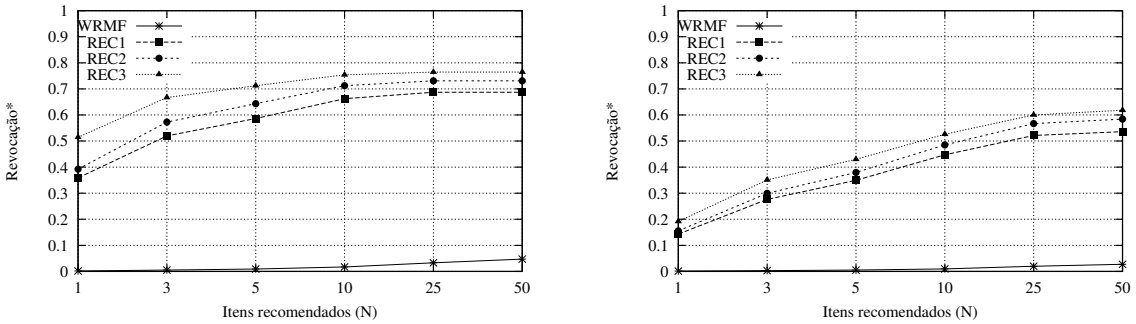
Figura 6.3. Resultados da aplicação da técnica de recomendação: precisão*.

A Figura 6.3 mostra que as técnicas propostas de recomendação obtiveram resultados de precisão* significantemente maiores que o método de *baseline*, alcançando valores na faixa de 70% a 90% para ambas as validações utilizadas.

Comparando-se as duas métricas de precisão abordadas nesse trabalho, é possível destacar a superioridade dos valores obtidos para o cálculo da métrica precisão*. Esse comportamento pode ser explicado pelo fato da aleatoriedade dos itens inseridos no *ranking* de recomendação, uma vez que existe uma vasta base de dados, que contém inúmeros objetos com *rankings* de valores muito baixos, o que favorece essa variação da métrica precisão.

6.5.5 Resultados para revocação*

Os resultados para revocação*, variação da métrica de revocação detalhada na Seção 6.4.2.2, são expostos na Figura6.4.



(a) Passado Recente - Futuro Recente

(b) Passado Recente - Futuro Completo

Figura 6.4. Resultados da aplicação da técnica de recomendação: revocação*.

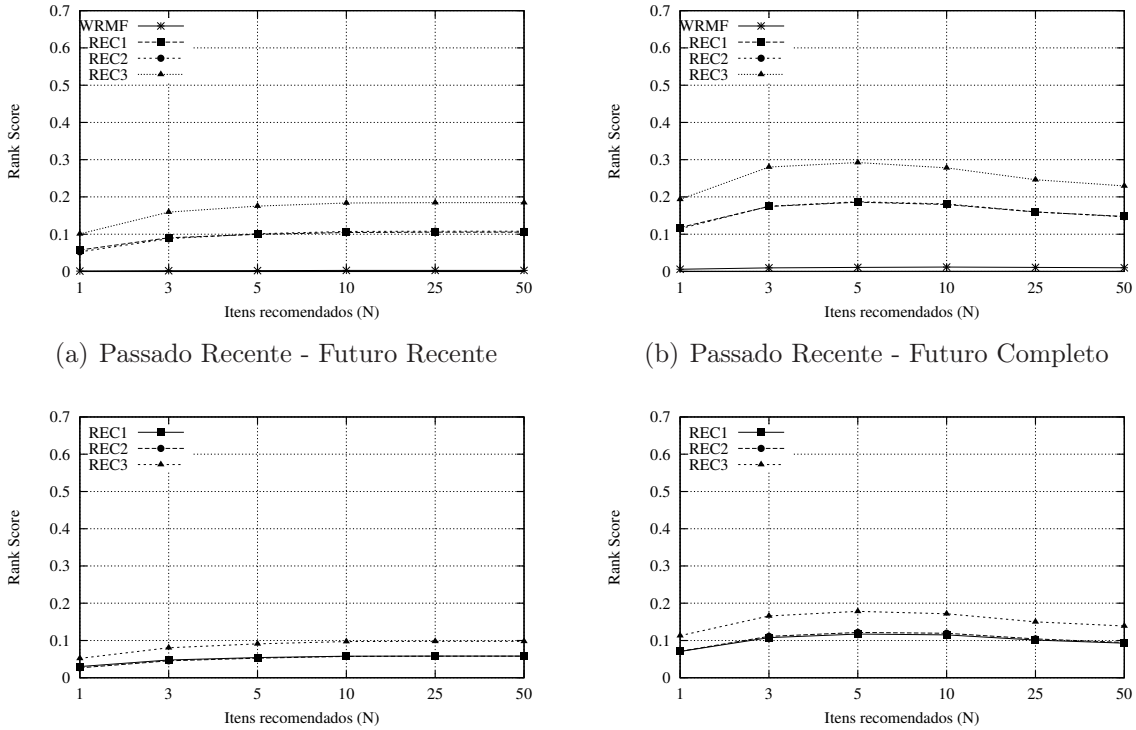
Novamente, as técnicas de recomendação com foco no objeto obtiveram melhores resultados do que o método WRMF (*baseline*), considerando a métrica de revocação*. Assim como os resultados para a métrica de revocação demonstrados na Seção 6.5.3, a validação utilizando apenas um vídeo visualizado pelo usuário (Passado Recente - Futuro Recente, Figura 6.4(a)) para a medida de eficiência obteve os melhores resultados em termos de revocação*, chegando a um valor de 76,46% para 50 itens recomendados utilizando a combinação das dimensões do objeto e de seu consumo (REC3).

6.5.6 Resultados para *Rank Score*

A seguir, estão os resultados da métrica de *Rank Score* (Seção 6.4.2.3), que considera a posição do item no *ranking* em seu cálculo.

A Figura 6.5 mostra que aumentar a quantidade de itens recomendados nem sempre é a melhor opção. Utilizando uma extensão da métrica de revocação, o *Rank Score* considera o posicionamento de um item corretamente recomendado para o seu cálculo, o que justifica os valores apresentados nessa figura.

Em geral, os resultados do *baseline* foram, mais uma vez, inferiores aos das recomendações propostas em nosso trabalho. A combinação das dimensões propostas por REC3 apresentou novamente os melhores valores de *Rank Score*, em destaque para a validação Passado Recente - Futuro Completo (Figura 6.5(b)), onde o valor dessa métrica para 5 itens foi igual a 29.24%. Esse comportamento pode ser explicado pelo fato de, nesse caso, serem utilizados mais itens para contabilizar o cálculo de *Rank Score*.



(a) Passado Recente - Futuro Recente

(b) Passado Recente - Futuro Completo

(c) Passado Completo - Futuro Recente

(d) Passado Completo - Futuro Completo

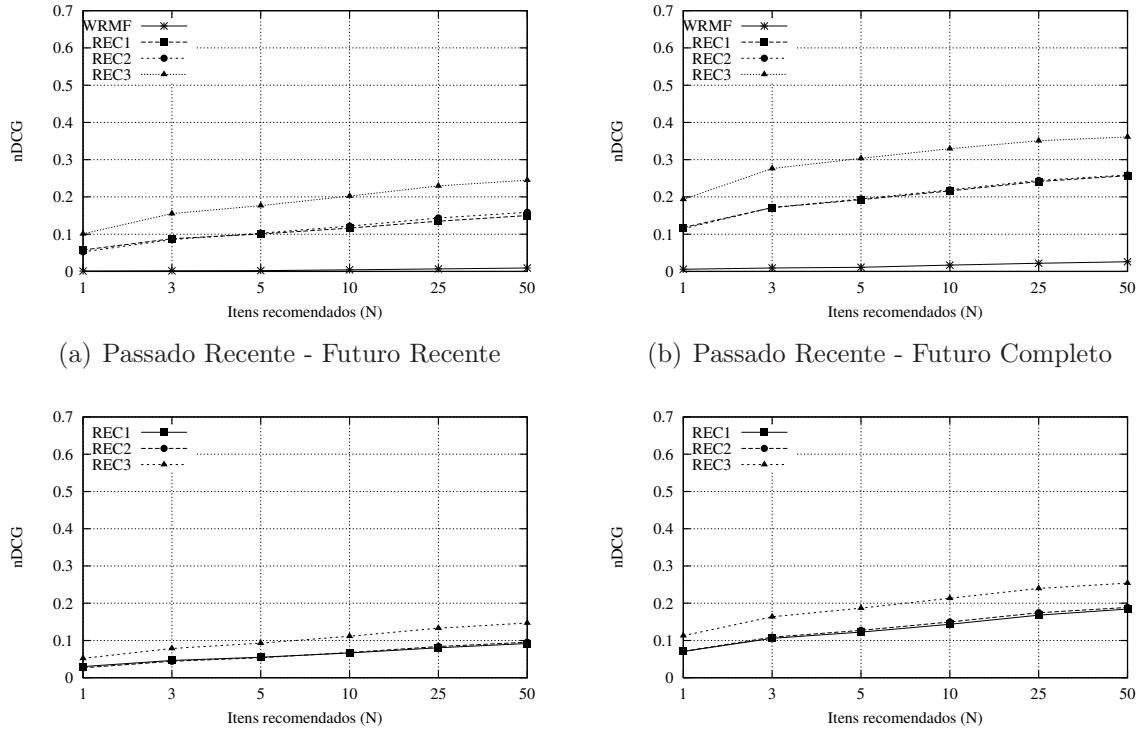
Figura 6.5. Resultados da aplicação da técnica de recomendação: *Rank Score*.

6.5.7 Resultados para nDCG

Apresentamos nesta seção os resultados de outra métrica que leva em consideração a posição no *ranking* de um item corretamente recomendado: o nDCG, ou *Normalized Discounted Cumulative Gain*, detalhado na Seção 6.4.2.4.

A Figura 6.6 demonstra que o aumento do número de itens recomendados resulta em uma melhora da métrica nDCG. Embora essa métrica considere o posicionamento da item no *ranking*, tal valor possui um peso logarítmico, e consequentemente, possui um impacto menor se comparado com a métrica de *Rank Score*.

Similarmente aos resultados do *Rank Score*, o nDCG apresentou maiores valores na validação de Passado Recente - Futuro Completo (Figura 6.6(b)), obtendo um valor de 36,10% ao recomendarmos 50 itens. Com valores próximos de zero, nosso *baseline* alcançou os piores resultados, e a utilização de dimensões do objeto para a recomendação (REC1) obteve, de maneira geral, valores iguais ao considerarmos apenas as dimensões do consumo do objeto para o mesmo propósito (REC2).



(a) Passado Recente - Futuro Recente

(b) Passado Recente - Futuro Completo

(c) Passado Completo - Futuro Recente

(d) Passado Completo - Futuro Completo

Figura 6.6. Resultados da aplicação da técnica de recomendação: nDCG.

6.5.8 Considerações Finais

Nas seções anteriores, foram apresentados os resultados da técnica de recomendação proposta aplicada na plataforma da Samba Tech. Com base em quatro possíveis validações, comparamos tais resultados com um *baseline* utilizando o método WRMF.

A partir das análises realizadas, podemos concluir, de uma maneira geral, o seguinte:

- Com o aumento do número de itens recomendados, a qualidade da recomendação apresenta melhoras na maioria dos casos;
- O método de *baseline* obteve os piores resultados em todos experimentos, mesmo sendo utilizado um algoritmo considerado estado-da-arte na literatura. Porém, é importante ressaltar que, como detalhado na Seção 6.3, utilizamos apenas a informação de visualização de objetos para a geração dos itens recomendados. A quantidade de dados utilizados como entrada é significativamente menor do que aquela considerada em nossas recomendações, o que impacta diretamente nos resultados obtidos;

- A utilização das dimensões do objeto e do seu consumo, quando aplicadas separadamente (REC1 e REC2, respectivamente), obtiveram em sua grande maioria, resultados similares, com uma leve superioridade de REC2;
- Ao combinar as dimensões do objeto e seu consumo (REC3), foi obtida a melhor eficácia de nossa técnica de recomendação e todos os experimentos.

Sendo assim, conclui-se que ao agregar mais informações que envolvem o objeto e o seu consumo, pode-se gerar uma lista de recomendação que, empiricamente, demonstrou-se ser de melhor qualidade para o usuário quando comparada com listas de recomendação que utilizem tais informações separadamente. Dessa forma, a hipótese proposta na Seção 6.1 foi validada:

> *Quanto mais dimensões do consumo do item utilizarmos para a comparação entre itens, mais refinada será a geração de nossos* rankings, *e consequentemente, faremos uma recomendação melhor.*

Por fim, com as análise realizadas e resultados expostos, também conclui-se que a recomendação de 5 itens seria a melhor opção para o recomendador, uma vez que este valor trouxe ganhos significativos com relação ao *baseline*. Além disso, obteve um valor de precisão superior em relação a N=1, e ainda sugerindo um conjunto de itens que é razoável no que diz respeito à recomendação, considerando que sugerir 5 itens é algo comumente utilizado em um cenário real.

Capítulo 7

Conclusão e Trabalhos Futuros

Neste trabalho, realizamos um estudo de serviços de conteúdo *Web* multimídia. Em meio a diversos cenários e aplicações, focamos tal análise no ambiente corporativo. Para isso, contamos com uma base de dados real disponibilizada pela Samba Tech, empresa que possui a maior plataforma de vídeos online corporativos da América Latina.

Para esse estudo, apresentamos uma abordagem com um foco diferente do usual: o objeto. Nesse cenário, temos como centro o conteúdo multimídia sendo consumido por um usuário. A partir de sua modelagem representativa, conseguimos identificar as entidades e responsabilidades que envolvem um serviço de conteúdo *Web* multimídia.

Com base nesse modelo, caracterizamos os dados fornecidos pela plataforma da Samba Tech, onde conseguimos compreender melhor o conteúdo que está sendo avaliado. O conhecimento adquirido a partir dessas análises foram importantes para o aprendizado da base de dados, e contribuíram para o direcionamento e tomada de decisão do projeto.

Propusemos um modelo e uma técnica de recomendação apoiada no foco do objeto que está sendo consumido. A modelagem nos permitiu definir três entidades: a primeira envolve o **objeto** ofertado ao usuário, assim como os seus metadados e especificações; as duas seguintes são referentes ao seu consumo, sendo divididas entre **como**/**quando**/**onde** um conteúdo é consumido e **quem** o consome. Além disso, o modelo proposto oferece diferentes focos de análise, de acordo com as entidades e objetivos envolvidos, que são utilizados com base para a técnica de recomendação elaborada. Ambientada em sistemas de recomendação do tipo *Top-N*, essa técnica propõe a geração de listas de potenciais itens (*rankings*) a partir da similaridade entre itens. Essa comparação é realizada com base em diferentes dimensões, escolhidas a partir das entidades existentes no modelo.

A técnica de recomendação proposta foi aplicada no cenário da plataforma da

Samba Tech a partir de três abordagens, que utilizavam dimensões apenas do objeto, dimensões apenas de como/quando/onde ele é consumido, e da combinação desses dois conjuntos. Além disso, realizamos a comparação dessa aplicação com um método considerado estado-da-arte em sistemas de recomendação do tipo *Top-N*: o WRMF (*Weighted Regularized Matrix Factorization*). Para os experimentos, propusemos também um método de validação para sistemas de recomendação, que foi utilizado em nossas análises em conjunto com diversas métricas de validação, como precisão, revocação e nDCG (*Normalized Discounted Cumulative Gain*).

Dentre os principais resultados obtidos, destacamos a superioridade de nossa técnica sobre o WRMF em todos os experimentos, embora a quantidade de informações utilizadas como entrada do algoritmo estado-da-arte tenha sido significantemente menor do que a utilizada em nossa técnica. Além disso, a recomendação que combina dimensões do objeto e do seu consumo alcançou uma precisão de quase 70% ao recomendarmos 50 vídeos. Esse valor chega a aproximadamente 40% quando recomendamos 5 itens, valor considerado, a partir de nossas análises, como ideal para ser utilizado na prática por um sistema de recomendação.

Os resultados obtidos demonstram a eficácia de nossa técnica de recomendação, mesmo em um cenário onde a recomendação de um objeto não é realizada de fato ao usuário. A análises também permitiram validar a hipótese proposta: *Quanto mais dimensões do consumo do item utilizarmos para a comparação entre itens, mais refinada será a geração de nossos* rankings, *e consequentemente, faremos uma recomendação melhor.*

Em suma, fomos capazes de modelar e caracterizar serviços de conteúdo *Web* multimídia, além de propormos um modelo e técnica de recomendação com foco no objeto, com a sua validação realizada utilizando dados reais. Com isso, temos como grandes contribuições dessa pesquisa o modelo representativo desse tipo de serviço, a modelagem e técnica de recomendação proposta, assim como o método de validação utilizado em sua análise. Esse resultados são de grande utilidade para provedores de conteúdo e usuários, uma vez que podem ser aplicados em diversos cenários de personalização de serviços, além da vasta aplicabilidade em sistemas de recomendação. Além disso, o método de validação proposto pode ser utilizado em inúmeras pesquisas que envolvam sistemas de recomendação.

Concluímos dessa maneira o estudo realizado nessa dissertação de mestrado, que também obteve como resultado duas submissões de artigos: no XIX Simpósio Brasileiro de Sistemas Multimídia e *Web* (WEBMEDIA'2013), tivemos a submissão e aceite de um artigo curto titulado *Modelagem, Caracterização e Recomendação em Serviços de Conteúdo Web Multimídia*. Também foi feita a submissão de um artigo para o evento

internacional ISM'2013 (*IEEE International Symposium on Multimedia*), intitulado *Modeling, Characterization and Recommendation of Multimedia* Web *Content Services*, que ainda está em fase de avaliação. Ainda será organizado um artigo para um periódico de boa qualidade na área de sistemas multimídia.

Como trabalhos futuros, esperamos estudar mais profundamente diferentes dimensões de consumo dos objetos para o cálculo de similaridade. Acreditamos que ainda tenhamos dimensões a serem estudadas e inferidas, além de podermos ponderar as dimensões do objeto. Além disso, pretendemos abordar mais a informação de quem consome o objeto. Por não termos nenhuma informação detalhada do usuário, tal tarefa se demonstrou um desafio para a pesquisa. Porém, acreditamos que, com a utilização de diferentes técnicas, podemos encontrar uma maneira de agregar esse tipo de informação à técnica de recomendação proposta. Por fim, vamos estudar a aplicação de diferentes métricas para a validação de nossa recomendação, o que permitirá realizarmos uma comparação de nossa técnica em diversos cenários e situações. Além disso, pretendemos, em parceria com a Samba Tech, aplicar essas técnicas em um cenário real, agregando valor tanto para a pesquisa quanto para a própria empresa.

Referências Bibliográficas

[And10] C. Anderson. *The Long Tail: How Endless Choice is Creating Unlimited Demand*. Random House, 2010.

[ASP00] Soam Acharya, Brian Smith, and Peter Parnes. Characterizing User Access To Videos On The World Wide Web. In *Proc. SPIE*, 2000.

[BPR+09] F. Benevenuto, A. Pereira, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Avaliação do perfil de acesso e navegação de usuários em ambientes web de compartilhamento de vídeos. In *Proceedings of the Simpósio Brasileiro de Sistemas Multimídia e Web (WEBMEDIA)*, Fortaleza, Brasil, 2009. SBC.

[BSS+08] Shumeet Baluja, Rohan Seth, D. Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 895–904, New York, NY, USA, 2008. ACM.

[CKR+09] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Trans. Netw.*, 17:1357–1370, October 2009.

[CKT10] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 39–46, New York, NY, USA, 2010. ACM.

[CLD13] X. Cheng, J. Liu, and C. Dale. Understanding the characteristics of internet short video sharing: A youtube-based measurement study. *Multimedia, IEEE Transactions on*, 15(5):1184–1194, 2013.

[CWVL01] Maureen Chesire, Alec Wolman, Geoffrey M. Voelker, and Henry M. Levy. Measurement and analysis of a streaming-media workload. In *Proceedings of the 3rd conference on USENIX Symposium on Internet Technologies and Systems - Volume 3*, USITS'01, pages 1–1, Berkeley, CA, USA, 2001. USENIX Association.

[DADSTN12] Ernesto Diaz-Aviles, Lucas Drumond, Lars Schmidt-Thieme, and Wolfgang Nejdl. Real-time top-n recommendation in social streams. In *Proceedings of the sixth ACM conference on Recommender systems*, RecSys '12, pages 59–66, New York, NY, USA, 2012. ACM.

[DK11] C. Desrosiers and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, chapter 4, pages 107–174. Springer US, Boston, MA, 2011.

[DLL+10] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 293–296, New York, NY, USA, 2010. ACM.

[GPnMG09] Roberto García, Xabiel G. Pañeda, David Melendi, and Victor Garcia. Probabilistic analysis and interdependence discovery in the user interactions of a video news on demand service. *Comput. Netw.*, 53:2038–2049, August 2009.

[GRFST11] Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. MyMediaLite: A free recommender system library. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011)*, 2011.

[GTD+11] Charles Gonçalves, Luam Totti, Diego Duarte, Wagner Meira Jr., and Adriano Pereira. Rock: Uma metodologia para a caracterização de serviços web multimídia baseada em hierarquia da informação. In *XVII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), 2011*, pages 174 – 181, Florianópolis, SC, 2011. Anais do XVII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), 2011.

[HKV08] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE*

International Conference on Data Mining, ICDM '08, pages 263–272, Washington, DC, USA, 2008. IEEE Computer Society.

[JK02] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[JZFF11] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender Systems An Introduction*. Cambridge University Press, 2011.

[LK04] Qing Li and ByeongMan Kim. Constructing user profiles for collaborative recommender system. In JeffreyXu Yu, Xuemin Lin, Hongjun Lu, and Yanchun Zhang, editors, *Advanced Web Technologies and Applications*, volume 3007 of *Lecture Notes in Computer Science*, pages 100–110. Springer Berlin Heidelberg, 2004.

[PZC+08] Rong Pan, Yunhong Zhou, Bin Cao, N.N. Liu, R. Lukose, M. Scholz, and Qiang Yang. One-class collaborative filtering. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 502–511, 2008.

[RRSK11] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[Sam04] Samba Tech. Samba Mobile Multimidia S/A. `http://www.sambatech.com`, 2004. [Online; acessado em 01-Janeiro-2011].

[SG11] Guy Shani and Asela Gunawardana. Evaluating Recommendation Systems Recommender Systems Handbook. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, chapter 8, pages 257–297. Springer US, Boston, MA, 2011.

[SK09] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.

[SKKR01] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[SMZ04] Kunwadee Sripanidkulchai, Bruce Maggs, and Hui Zhang. An analysis of live streaming workloads on the internet. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, IMC '04, pages 41–54, New York, NY, USA, 2004. ACM.

[SPUP02] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 253–260, New York, NY, USA, 2002. ACM.

[TSK05] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[VAM+02] Eveline Veloso, Virgílio Almeida, Wagner Meira, Azer Bestavros, and Shudong Jin. A hierarchical characterization of a live streaming media workload. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, IMW '02, pages 117–130, New York, NY, USA, 2002. ACM.

[Zip32] G. K. Zipf. Selective Studies and the Principle of Relative Frequency in Language, 1932.