Sistema para Recomendações de Animes por meio de Redes Heterogêneas e Aprendizado Transdutivo

Brucce N. dos Santos[1], Julio César Carnevali[1], Luan V. de C. Martins[1], Michael J. Bianchi[2]

[1]ICMC – Universidade de São Paulo (USP)
Caixa Postal 359 – 13566-590 – São Carlos – SP – Brazil

[2]EESC – Universidade de São Paulo (USP)
Caixa Postal 359 – 13566-590 – São Carlos – SP – Brazil

brucce.neves@gmail.com,{carnevali,luan.martins,michael_bianchi}@usp.br

Abstract. *Due to the popularization of the Japanese culture, animes – Japanese movies made from animations – gained more notoriety. In this work, a content filtering based recommendation system through transductive learning in heterogeneous networks is proposed, using the textual description and the genres to recommend new items similar to the users' tastes. In this study, after analyzing the main techniques and choosing the GNetMine algorithm, we could conclude that the proposed model is capable of generating new acceptable recommendations and with low processing costs, even with very little training. For tests using a positively labeled and negatively labeled anime, the best result generated was 41% hit.*

Resumo. *Devido a popularização da cultura japonesa, animes – filmes de animações japonesas – ganharam mais notoriedade. Nesse trabalho, é proposto um sistema de recomendação de animes baseado em conteúdo através de aprendizado transdutivo em redes heterogêneas, usando a descrição textual e os gêneros para recomendar novos itens semelhantes aos gostos dos usuários. Nesse estudo, após análise das principais técnicas e escolha do algoritmo GNetMine, pudemos concluir que o modelo proposto é capaz de gerar novas recomendações e com baixo custo de processamento, mesmo com pouco treinamento. Para testes utilizando um anime rotulado positivamente e um rotulado negativamente, o melhor resultado gerado foi de 41% de acerto.*

1. Introdução

Toda forma de entretenimento permite ao ser humano renovar sua energia e humor para enfrentar as tarefas do cotidiano, permitindo viver com mais entusiasmo. Existem várias formas de entretenimento, desde músicas, filmes, animes, novelas, desenhos animados, entre outros.

Dentre esses, os animes se destacam, apresentando um crescimento significativo nos últimos anos. Segundo o último relatório da *Association of Japanese Animation* (AJA)[1], o mercado de animes registrou aumento nas vendas por 4 anos consecutivos, enquanto o tamanho do mercado ultrapassou a marca de 2 trilhões de ienes (aproximadamente 17.6 bilhões de dólares), impulsionado pelas vendas no exterior. Grande parte

[1]http://aja.gr.jp/english/japan-anime-data Acessado em 02/12/2018

desse valor é correspondente ao mercado internacional, o qual registrou um crescimento de 171,9% desde 2013.

Impulsionada por esse crescimento, a quantidade de animes disponíveis no mercado aumenta a cada dia. Com isso, a pesquisa para encontrar algo que mais nos agrada exige cada vez mais tempo, o qual o indivíduo não está disposto a desperdiçar com atividades que não agregam valor para seu dia. Mecanismos capazes de realizar recomendações de itens para os usuários, diminuindo a necessidade de intervenção humana, tem ganhado importância nas últimas décadas [Rossi et al. 2015].

Os denominados sistemas de recomendação auxiliam nessa questão, sendo capazes de analisar as preferências de um determinado usuário e direcionar sugestões consideradas interessantes dentre todos os dados disponíveis, reduzindo o trabalho do usuário. Um dos grandes desafios deste tipo de sistema é realizar o casamento correto entre o que está sendo recomendando e aqueles que estão recebendo a recomendação, ou seja, definir e descobrir este relacionamento de interesses [Reategui and Cazella 2005].

Nesse estudo, é proposto um sistema de recomendação de animes baseado em conteúdo através de aprendizado transdutivo em redes heterogêneas. Com o objetivo de melhorar os resultados e minimizar o fenômeno de "especialização excessiva", frequentemente encontrado em sistemas do tipo, duas diferentes visões são utilizadas para a geração da rede: a descrição textual e o gênero do anime.

Para isso, a classificação automática de textos demonstra ser interessante, a qual consiste em atribuir rótulos automaticamente para os dados contidos nos textos selecionados para análise, possibilitando encontrar padrões na coleção de documentos analisados. Uma forma de realizar a classificação automática de textos é representar coleções de documentos por meio das chamadas redes [Blanco and Lioma 2012].

Existem diversos métodos para geração de redes e algoritmos de classificação de textos baseados em redes, dentre eles o uso de algoritmos de aprendizado transdutivo, como o **GFHF** - *Gaussian Fields and Harmonic Functions* [Zhu et al. 2003], **LLGC** - *Learning With Local and Global Consistency* [Zhou et al. 2004], **LPHN** - *Label Propagation through Heterogeneus Networks* [Rossi et al. 2014] e o **GNetMine** [Ji et al. 2010].

Essa estratégia (aprendizado transdutivo) foi considerada no presente estudo a fim de recomendar animes com semelhança de conteúdo (descrição) e de gênero, permitindo a recomendação de novos títulos para o usuário, uma vez que dois ou mais animes podem pertencer ao mesmo gênero, porém apresentarem descrições e contextos muito diferentes entre si.

1.1. Objetivos

O presente estudo visa investigar o uso de aprendizado transdutivo em redes, a fim de propor um sistema de recomendação de animes por meio de redes heterogêneas, que permita considerar diferentes tipos de objetos e relações a fim de fornecer recomendações confiáveis para o usuário.

O sistema de recomendação proposto visa fornecer ao usuário recomendações de animes que estejam alinhadas com as suas preferências, considerando a experiência e gostos desse usuário. Com isso é possível prever (com certo grau de confiança) quais os próximos animes que o usuário se interessaria em assistir, permitindo recomendar

um conteúdo alinhado com as suas necessidades. Para tal, foi utilizado o algoritmo de aprendizado transdutivo conhecido como GNetMine [Ji et al. 2010], a fim de atender os objetivos do estudo.

1.2. Organização do Texto

O restante deste artigo está organizado da seguinte forma. Na Seção 2 são apresentados os conceitos básicos sobre sistemas de recomendação e aprendizado transdutivo em redes. A Seção 3 traz detalhes da modelagem do problema. A seção 4 apresenta os resultados obtidos com o estudo. Finalmente, a Seção 5 apresenta as conclusões e perspectivas futuras.

2. Fundamentos

Neste capítulo é apresentado o referencial teórico que fundamenta o presente estudo, de forma que o leitor se familiarize com o tema. Os assuntos abordados são: os principais conceitos sobre Sistemas de Recomendação e Aprendizado Transdutivo em Redes Heterogêneas.

2.1. Sistemas de Recomendação

Sistemas de recomendação podem ser entendidos como técnicas e ferramentas de software capazes de gerar sugestões para itens com maior probabilidade de interesse para um usuário em específico [Resnick and Varian 1997, Burke 2007]. O usuário menciona um conjunto de palavras ou indicadores e os itens identificados como os mais relevantes são recomendados por algoritmos de busca e recuperação [Motta et al. 2011]. A palavra "Item" nesse caso é usada para denotar o que o sistema recomenda para o usuário [Ricci et al. 2015], podendo ser páginas web, filmes, músicas, livros, artigos, etc.

Em sua forma mais simples, recomendações personalizadas são oferecidas como listas classificadas de itens, a fim de prever quais são os produtos ou serviços mais adequados baseados nas preferências e restrições do usuário [Ricci et al. 2015]. Um objetivo essencial dos sistemas de recomendação é ajudar os usuários a realizarem melhores escolhas [Jameson et al. 2015].

Existem diferentes tipos de sistemas de recomendação que variam em termos do domínio abordado e do conhecimento utilizado, mas especialmente no que diz respeito ao algoritmo de recomendação, ou seja, como é feita a previsão da utilidade de um item [Ricci et al. 2015]. Entre os tipos existentes, os mais comuns são:

Recomendação baseada em Conteúdo: A Filtragem por Conteúdo usa algoritmos de aprendizado de máquina para induzir um perfil das preferências de um usuário a partir de exemplos, tendo em vista uma descrição das características dos conteúdos.Nesse caso, o sistema aprende a recomendar itens semelhantes aos que o usuário gostou no passado [Ricci et al. 2015]. Por exemplo, se um usuário classificou positivamente um filme que pertence ao gênero de ação, o sistema pode aprender a recomendar outros filmes do mesmo gênero. Dentre as vantagens dessa abordagem estão o fácil entendimento por parte do usuário e o baixo custo de aplicação, entretanto apresenta especialização excessiva, onde o usuário está restrito a visualizar itens semelhantes aos que já foram experimentados no passado [Soares and Viana 2015].

Recomendação baseada em Filtragem Colaborativa: As recomendações nesse caso baseiam-se na colaboração entre os grupos de interessados, ou seja, em itens que outros usuários com gostos semelhantes ao usuário ativo gostaram no passado. Dentre as vantagens dessa abordagem estão a facilidade de lidar com qualquer tipo de conteúdo e a recomendação de itens com conteúdo diferente daqueles experimentados anteriormente, entretanto a falta de dados de preferência do usuário causa um declínio de desempenho e dificulta a localização de vizinhos mais próximos para usuários com preferências específicas, além de ser um método considerado computacionalmente caro [Soares and Viana 2015].

Sistemas Híbridos de Recomendação: A abordagem híbrida se baseia na combinação da filtragem colaborativa e da filtragem baseada em conteúdo visando utilizar as vantagens de uma e corrigir as desvantagens da outra [Herlocker et al. 2000].

2.2. Aprendizado Transdutivo em Redes Heterogêneas

Algoritmos de aprendizado de máquina visam aprender, generalizar ou encontrar padrões a partir de uma coleção de informações sobre um determinado problema. Esses algoritmos são comumente organizados em três categorias: supervisionado, semissupervisionado e não supervisionado [Mitchell 1997].

O foco desse trabalho é no aprendizado semissupervisionado, pois nesse paradigma são utilizados tanto exemplos não rotulados quanto exemplos rotulados durante a tarefa de aprendizado [Zhu and Goldberg 2009a]. O objetivo é aumentar o conjunto de exemplos disponíveis para treinamento, visando melhor discriminar os padrões [Chapelle et al. 2006]. Em geral, esse tipo de aprendizado costuma ser utilizado quando existem poucos exemplos rotulados, sendo útil em diversas aplicações práticas, uma vez que normalmente é fácil coletar exemplos, mas é custoso rotulá-los devido à necessidade de um especialista do domínio e o tempo gasto para rotulação.

Os algoritmos de aprendizado semissupervisionado podem ser divididos em algoritmos de aprendizado transdutivo e algoritmos de aprendizado indutivo. No aprendizado transdutivo não existe a necessidade de construir um modelo de classificação uma vez que os exemplos não rotulados são classificados diretamente durante o próprio processo de aprendizado, sem necessidade de um estágio para treino e para teste [Gammerman et al. 1998]. O algoritmo é capaz de observar todos os exemplos a serem classificados, explorando suas características e consequentemente ir melhorando a performance de classificação. Por outro lado, no aprendizado indutivo, os exemplos previamente rotulados juntamente com os exemplos não rotulados são usados para construir um modelo de classificação, a fim de classificar exemplos não vistos [Blum and Mitchell 1998].

Nesse trabalho estaremos utilizando os algoritmos de aprendizado transdutivo, que inicialmente foram propostos usando o modelo espaço-vetorial, entretanto, estudos anteriores reportaram que não foram obtidos resultados satisfatórios com esse tipo de representação [Zhu and Goldberg 2009b]. Por outro lado, estudos recentes têm apresentado resultados promissores de classificação transdutiva por meio de redes [Rossi et al. 2015].

O uso de redes (representação relacional) é uma alternativa ao modelo espaço-vetorial para representação dos dados textuais, que vem ganhando atenção nas últimas

décadas [Sun and Han 2012, Silva and Zhao 2012, Newman 2010]. A maior vantagem dessa representação é tornar explícito os diferentes tipos de relações entre os diferentes tipos de entidades dos textos [Rossi et al. 2016]. Estudos recentes indicam que a classificação de textos por meio de redes alcança resultados promissores quando comparados com o modelo espaço-vetorial [Breve et al. 2012, Rossi et al. 2015, Rossi et al. 2016].

Uma rede é formalmente definida como uma tripla $\mathcal{N} = \langle \mathcal{O}, \mathcal{R}, \mathcal{W} \rangle$, em que $\mathcal{O}$ representa o conjunto de objetos (ou elementos) da rede, $\mathcal{R}$ representa o conjunto das relações entre os objetos e $\mathcal{W}$ representa o conjunto de pesos das relações. Se o conjunto $\mathcal{O}$ possuir um único tipo de objeto, esse tipo de rede é denominada rede homogênea. Se o conjunto $\mathcal{O}$ possui dois ou mais tipos de objetos é denominada rede heterogênea [Ji et al. 2010]. Dessa forma, consegue-se identificar relações entre as entidades de um problema de maneira eficiente, além de apresentar uma melhor performance de classificação [Breve et al. 2012].

Ao utilizar a modelagem de rede na tarefa de classificação, cada exemplo do conjunto de objetos $\mathcal{O}$ de uma rede possui um vetor de informação de classes $\mathbf{f}$. Cada posição do vetor armazena o nível de pertinência do exemplo a um dos rótulos. Caso seja um exemplo rotulado, ele possuirá um vetor $\mathbf{y}$ de mesmas dimensões de $\mathbf{f}$, mas com o objetivo de armazenar o(s) rótulo(s) original(is) do exemplo. Nesse último caso, o vetor $\mathbf{y}$ é de tal forma que possui o valor 1 para determinar (no índice apropriado do vetor) em qual classe o exemplo está rotulado.

A classificação transdutiva em redes pode ser realizada considerando a estratégia de regularização, a qual tem a premissa de minimizar uma função objetivo satisfazendo duas condições: (i) as informações de classe de objetos vizinhos devem ser semelhantes; (ii) as informações de classe dos objetos rotulados atribuídas durante o processo de classificação devem ser semelhantes às informações de classe reais [Zhu 2005]. O framework para satisfazer essas duas condições está expresso na Equação 1 [Delalleau et al. 2005]:

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{o_i, o_j \in \mathcal{O}} w_{o_i,o_j} \Omega(\mathbf{f}_{o_i}, \mathbf{f}_{o_j}) + \mu \sum_{o_i \in \mathcal{O}^L} \Omega'(\mathbf{f}_{o_i}, \mathbf{y}_{o_i}) \quad (1)$$

O primeiro termo $\Omega(.)$ da função de regularização é responsável por computar a proximidade entre os vetores de informação de classe entre cada par de objetos relacionados na rede. Tal proximidade pode ser derivada de funções de distâncias ou de dissimilaridades. Já o segundo termo $\Omega'(.)$ é responsável por computar a proximidade entre a informação de classe de objetos rotulados e as respectivas informações reais de classe. Ainda nesta equação, w_{o_i,o_j} indica o peso da relação entre os objetos e o parâmetro μ indica a importância da informação de classe real durante a propagação dos rótulos.

Essa equação pode ser resolvida por meio de soluções iterativas, denominadas "propagação de rótulos" [Wang and Zhang 2006], onde os objetos propagam suas informações de classe (rótulos) para os objetos vizinhos de forma proporcional ao peso de suas relações. Nessas iterações, a informação de classe $\mathbf{f}$ de cada objeto é inicializada com valores zero (ou de forma aleatória). O critério de parada pode ser obtido pela convergência, ou seja, quando não há alterações significativas nas informações de classe dos

objetos. Outro critério de parada pode ser o número máximo de iterações.

A seguir são apresentados os principais algoritmos de classificação transdutiva em redes homogêneas e heterogêneas. Esses foram selecionados por serem os mais utilizados nesta área e representam as estratégias que são base para várias outras propostas [Zhu and Goldberg 2009a].

Um dos principais algoritmos para classificação transdutiva em redes homogêneas é o GFHF (***Gaussian Fields and Harmonic Functions***) [Zhu et al. 2003]. Sua principal característica é computar a informação de classe de um objeto não rotulado com base na média de informação de classe dos objetos vizinhos ponderada pelos pesos das respectivas conexões.

Uma adaptação do algoritmo GFHF foi proposta por [Zhou et al. 2004], denominado LLGC (***Learning With Local and Global Consistency***) também para redes homogêneas. No LLGC, é assumido que alguns exemplos podem ser erroneamente rotulados pelo usuário, causando com isso uma deterioração da performance de classificação. Dessa forma, o algoritmo permite que os exemplos rotulados possam ter seus rótulos alterados durante a classificação. Além disso, foi proposta modificação na força de propagação, em que se considera tanto o grau do objeto que está propagando quanto o grau do objeto que está recebendo a informação.

O algoritmo LPHN (***Label Propagation through Heterogeneus Networks***) [Rossi et al. 2014] é uma extensão do algoritmo GFHF para redes heterogêneas. Assim, é mantida a propriedade de que objetos rotulados pelos usuários não devem ter seus rótulos alterados. O LPHN usa uma função de regularização análoga à do GFHF, com a restrição de que a propagação considera diferentes tipos de relações.

Um algoritmo para propagação de rótulos em redes heterogêneas que considera a importância das relações, além de permitir reduzir a importância de objetos inicialmente rotulados foi proposto por [Ji et al. 2010], denominado GNetMine. Devido a possibilidade de definir a importância de cada relação ele foi utilizado como regularizador desse trabalho.

2.2.1. GNetMine

Na prática, o GNetMine pode ser visto como uma extensão do LLGC para redes heterogêneas, uma vez que o primeiro termo do regularizador também ameniza o efeito de objetos com um alto valor de grau. Apesar da possibilidade de definir a importância das relações entre os objetos, esse valor de importância deve ser definido previamente pelo usuário ou por algum processo de calibragem.

O GNetMine visa satisfazer as seguintes premissas: (I) a atribuição de classes a dois objetos relacionados tende a ser similar, e (II) a predição das classes de objetos rotulados tendem a ser similar as classes pré-definidas. A função de regularização do GNetMine é descrita na Equação 2.

$$Q(\mathbf{F}) = \sum_{\mathcal{O}_i,\mathcal{O}_j \subset \mathcal{O}} \lambda_{\mathcal{O}_i,\mathcal{O}_j} \sum_{o_k \in \mathcal{O}_i} \sum_{o_l \in \mathcal{O}_j} w_{o_k,o_l} \left\| \frac{\mathbf{f}_{o_k}(\mathcal{O}_i)}{\sqrt{\sum_{o_m \in \mathcal{O}_j} w_{o_k,o_m}}} - \frac{\mathbf{f}_{o_l}(\mathcal{O}_j)}{\sqrt{\sum_{o_m \in \mathcal{O}_i} w_{o_l,o_m}}} \right\|^2 + \sum_{o_j \in \mathcal{O}^L} \alpha_{o_j} (\mathbf{f}_{o_j} - \mathbf{y}_{o_j}) \quad (2)$$

Nesta função, $\lambda_{\mathcal{O}_i,\mathcal{O}_j}$ é utilizado para definir a importância da relação entre um objeto do tipo $\mathcal{O}_i$ e um objeto do tipo $\mathcal{O}_j$, onde $0 \leq \lambda_{o_i,o_j} \leq 1$ e o α_{o_j} possuindo valores entre $0 \leq \alpha_{o_j} \leq 1$. Já $\alpha_{o_j} \in \mathcal{O}^L$ define a importância (parâmetro α entre $[0, 1]$) de objeto inicialmente rotulado o_j. Abaixo é apresentado o algoritmo GNetMine.

Algoritmo 1: *GNetMine*

Entrada: $\mathcal{O}, \mathbf{Y}, \mathbf{W}, \mathbf{D}, \alpha, \lambda[]$
Saída: $\mathbf{F}(\mathcal{O}^U)$

1 **início**
2 **para** *cada* $\mathcal{O}_i, \mathcal{O}_j \subset \mathcal{O}$ **faça**
3 $\mathbf{S}(\mathcal{O}_i, \mathcal{O}_j) \leftarrow \mathbf{D}(\mathcal{O}_i, \mathcal{O}_j)^{-\frac{1}{2}} \cdot \mathbf{W}(\mathcal{O}_i, \mathcal{O}_j) \cdot \mathbf{D}(\mathcal{O}_i, \mathcal{O}_j)^{-\frac{1}{2}}$
4 **fim**
5 **fim**
6 **repita**
7 **para** *cada* $\mathcal{O}_i \subset \mathcal{O}$ **faça**
8 **para** *cada* $o_j \subset \mathcal{O}_i$ **faça**
9 $\mathbf{f}_{o_j} \leftarrow \sum_{\mathcal{O}_k \subset \mathcal{O}, o_j \notin \mathcal{O}_k} \lambda_{\mathcal{O}_i,\mathcal{O}_k} \cdot \sum_{o_m \in \mathcal{O}_k} S(\mathcal{O}_i, \mathcal{O}_k)_{o_j,o_m} \cdot \mathbf{f}_{o_m}$
10 $\mathbf{f}_{o_j} \leftarrow \mathbf{f}_{o_j} + \left(2 \cdot \lambda_{\mathcal{O}_i,\mathcal{O}_i} \cdot \sum_{o_m \in \mathcal{O}_i} S(\mathcal{O}_i, \mathcal{O}_i)_{o_j,o_m} \cdot \mathbf{f}_{o_m}\right) + \alpha_{o_j} \cdot \mathbf{y}_{o_j}$
11 $\mathbf{f}_{o_j} \leftarrow \frac{\mathbf{f}_{o_j}}{\sum_{\mathcal{O}_k \subset \mathcal{O}, o_j \notin \mathcal{O}_k} \lambda_{\mathcal{O}_i,\mathcal{O}_k} + 2 \cdot \lambda_{\mathcal{O}_i,\mathcal{O}_i} + \alpha_{o_j}}$
12 **fim**
13 **fim**
14 **até** *convergência ou outra condição de parada*;
15 **retorna** $\mathbf{F}(\mathcal{O}^U)$

A entrada do algoritmo é o conjunto de objetos da rede $\mathcal{O}$, as informações de classe reais $\mathbf{Y}$, os pesos de conexões entre objetos $\mathbf{W}$, grau dos objetos $\mathbf{D}$, a importância α das informações previamente rotuladas e um vetor $\lambda[]$ com a importância de cada tipo de relação. Na Linha 3 é definida um valor de confiança da conexão entre dois objetos na rede, dada dois tipos de relações. A propagação de rótulos é realizada nas Linhas $9 - 11$, em que é realizada a atualização de informação de classe de um objeto $o_j \in \mathcal{O}_i$ conforme os objetos vizinhos, a importância da relação dos objetos vizinhos e a importância da informação rotulada.

3. Modelagem do problema

O domínio de animes foi escolhido devido ao crescimento significativo do mercado nos últimos anos, o qual ultrapassou a marca de 2 trilhões de ienes segundo o último relatório da *Association of Japanese Animation* (AJA), conforme mencionado anteriormente. Devido a grande quantidade de animes disponíveis, os usuários tendem a ter dificuldade

para encontrar quais são de seu interesse, levando até à desistência por conta da dificuldade enfrentada durante a procura. Por motivos como este, sistemas de recomendação são demandados, facilitando a vida de quem passava horas a procura de algo, independente do meio inserido (animes, filmes, produtos, etc.).

Dentre os tipos de sistemas de recomendação, o presente estudo foca na recomendação baseada em conteúdo, uma vez que o usuário irá avaliar positivamente ou negativamente um conjunto de animes e os mais similares de acordo com suas preferências serão recomendados, note entretanto, que essa técnica possui o problema de especialização excessiva como citado na Seção 2.1. A fim de minimizar esse problema, o presente estudo propõe a criação de um sistema de recomendação baseado em aprendizado transdutivo em redes heterogêneas, permitindo rotular animes e gêneros, algo que não seria possível de ser realizado no espaço vetorial, já que o mesmo só permite rotular animes. A proposta é fornecer um sistema capaz de gerar sugestões de animes com maior probabilidade de interesse para o usuário e com custos operacionais mais baixos, quando comparado com a abordagem colaborativa.

3.1. Estrutura da Rede Proposta

O problema foi modelado em uma rede não dirigida usando três camadas conforme ilustrado na Figura 1, sendo a camada de "Animes" a camada alvo da recomendação, onde cada nó representa um anime e dois animes estarão conectados, caso eles pertençam a uma mesma franquia. A camada de "Gêneros" é composta por todos os gêneros presentes na coleção, possuindo conexões somente com nós da camada anime. A camada "Descrição" é a camada com a maior quantidade de nós na rede, cada nó representa uma palavra presente na descrição dos animes após o pré-processamento da descrição, sendo que cada nó está relacionado somente aos animes que possuem aquela palavra na descrição.

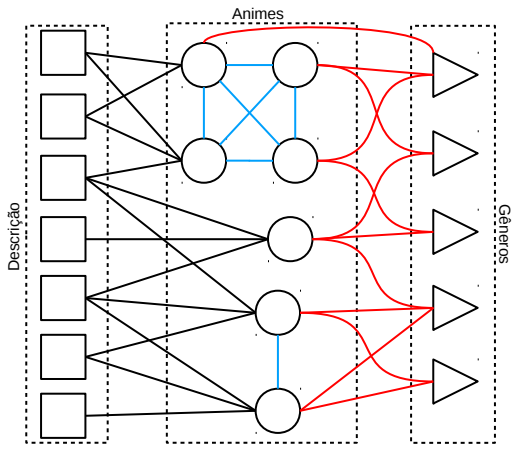

Figura 1. Ilustração de uma rede de animes.

A partir da avaliação positiva ou negativa do usuário sobre um determinado anime, o sistema irá identificar a qual gênero este se refere, buscando novos animes relacionados ao mesmo gênero, além de animes que possuam descrição semelhante ao anime avaliado positivamente, retornando assim, sugestões de animes coerentes com as preferências do usuário.

3.2. Fluxo de execução

Para fazer uso do sistema, o usuário deve providenciar o mínimo de treinamento para o algoritmo, rotulando pelo menos 1 gênero ou item que goste e 1 gênero ou item que não goste. Em seguida, a rede é normalizada por meio do algoritmo GNetMine, e os rótulos "gostou" e "não gostou" serão propagados pela rede, dando a cada anime um nível de pertinência a cada rótulo. Os animes que tiverem um grau de pertinência para a classe "gostou" maior que um limiar pré definido, serão então retornados como animes recomendados.

Com o objetivo obter melhores resultados, o sistema proposto também retornará ao usuário uma lista de animes no qual ele ficou "indeciso", ou seja, seu grau de pertinência para o rótulo "gostou" e para o rótulo "não gostou" são muito semelhantes. Dessa forma, quando o usuário fornecer sua opinião sobre esses itens e executar o algoritmo novamente, ele terá resultados mais precisos do que os encontrados em execuções anteriores – característica que recebe o nome de "active learning" [Settles 2012].

4. Avaliação experimental

Nesta seção é apresentada a base de dados que fora utilizada para a avaliação do método proposto, bem como a configuração experimental, os critérios de avaliação que foram utilizados e finalmente os resultados alcançados.

4.1. Descrição da Coleção de Animes

A coleção consiste em 7403 animes coletados do site MyAnimeList[2] por meio de *web crawling*. Dentre as informações presentes em cada página dos animes do *site*, foram coletadas o título, gêneros, descrição, data do lançamento, duração dos episódios, a idade mínima recomendada e o estúdio que o produziu. A descrição dos animes está na língua inglesa e contém algumas palavras em *Romanji*[3]. Nem todas as informações coletadas foram utilizadas na rede, sendo que a rede é montada com a descrição, os gêneros e o nome dos animes. No entanto, as outras informações foram úteis na parte gráfica de interação com o usuário.

Durante a atividade de pré-processamento dessa base, foram removidos os animes de conteúdo adulto, e foram apagadas as descrições que eram muito pequenas (menor que 50 caracteres), pois estas não continham informações que descreviam o anime. Por exemplo "*Specials included on the DVDs*".

Logo em seguida, as *stop-words*, palavras que não adicionam significado ao texto, tal como "the" ou "or", da língua inglesa foram removidas. Para isso foi utilizada a biblioteca *nltk.corpus.stopwords* disponível para Python. Também foram retiradas pontuações como " ? ! . , - : ; ". *Stop-words* de domínio também foram removidas, as quais não eram relevantes para a descrição do anime, por exemplo: "[*Written by* ...]", "(*Source*: ...)".

Feito isso, foi gerada a *bag-of-words* (BoW) de todas as descrições, sendo utilizada a medida TF-IDF (*Term Frequency - Inverse Document Frequency*), frequentemente utilizada para esta atividade de geração da BoW. Especificamente para este trabalho, os

[2]https://myanimelist.net/, Acessado em 18/11/2018

[3]Transcrição fonética dos caracteres japoneses para o alfabeto latino

valores da BoW foram normalizados para um intervalo entre 0 e 1, para que quando fossem geradas as arestas entre animes e termos, os pesos (que são os valores da BoW) estivessem todos dentro do mesmo intervalo.

A rede foi então gerada a partir deste *corpus* pré-processado. As conexões entre animes foram feitas com base nos itens que fazem parte da mesma franquia (sequências, *spin-offs*, *prequels*, filmes, etc.) sempre com peso constante 1. Já as conexões entre animes e os termos uilizados em suas descrições, foram geradas com base no valor TF-IDF obtido da BoW: a aresta é gerada com o peso do valor, sempre que o *TF-IDF* em questão é maior que 0. Por fim, as conexões entre animes e gêneros foram criadas com base nos gêneros aos quais o anime faz parte, tendo as arestas peso constante 1.

A distribuição dos nós por camada é mostrada na Tabela 1. Na camada "Animes" são encontrados 7403 vértices, onde cada um é um anime. Já na camada "Descrição", existem 30999 vértices, onde cada vértice representa um termo da *bag-of-words*. Finalmente, na camada "Gêneros", estão 42 vértices, sendo cada um, um gênero distinto.

Camada	Quantidade de nós
Animes	7403
Descrição	30999
Gêneros	42

Tabela 1. Descrição da rede.

4.2. Configuração Experimental

Devido a complexidade de avaliar um sistema de recomendação e a limitação de tempo, a avaliação experimental foi realizada utilizando as preferências de 5 usuários, sendo que cada usuário selecionou um conjunto de animes que já assistiu e os classificou como "gostou" e "não gostou", na Tabela 2 é apresentado a distribuição dos animes rotulados pelos usuários.

	Animes Rotulados		
Usuários	Gostou	Não Gostou	Total de Pares
Brucce	102	85	8670
Julio	11	5	55
Luan	27	13	351
Michael	13	14	182
Toshio	26	14	364

Tabela 2. Distribuição dos animes rotulados por usuário.

Também devido a limitação de tempo não foi possível avaliar o impacto na recomendação ao usar GNetMine com diferentes combinações de parâmetros, dessa forma o experimento foi executado com uma única configuração sendo o nível μ de importância da informação rotulada 0.75, além disso, o GNetMine permite definir a importância de cada relação de uma rede heterogênea, a rede proposta nesse trabalho possui 3 tipos de relações onde a relação entre animes da mesma franquia teve a importância definida como 0.45, a relação entre animes e a camada descrição recebeu a importância de

0.45 e por fim, a importância da relação entre animes e a camada de gêneros foi definida como 0.1.

O processo de avaliação consistiu em selecionar um par de animes rotulados e usar eles no processo de propagação de rótulos, sendo que cada par é composto por um anime rotulado como "gostou" e outro rotulado como "não gostou". Após a rede convergir, foi avaliado as recomendações nos critérios definidos na Seção 4.3. Esse processo foi repetido para todos os possíveis pares de cada conjunto de animes rotulados pelo usuário.

4.3. Critérios de Avaliação e Resultados

Foram utilizados 2 critérios de avaliação: o primeiro critério consiste em ordenar os resultados com base na pertinência deles a classe positiva (anime recomendado), sendo que quanto mais perto da primeira posição, maior é a pertinência desse anime a essa classe. Após isso, é verificado se o anime mais bem ranqueado pertence ao conjunto de animes rotulados por aquele usuário, caso pertença, então o par de animes utilizados para gerar aquela recomendação acertou. Após avaliar todos os pares nesse critério, a taxa de acertos é calculada. Esse critério é muito rígido uma vez que só considera acerto quando um anime do conjunto rotulado está na primeira posição.

Os resultados obtidos com esse critério de avaliação estão ilustrados na Figura 2, onde é possível observar que, em grande parte dos casos de acertos nesse critério, o anime melhor ranqueado pertence a mesma franquia ou é muito semelhante ao anime rotulado como "gostou", ao mesmo tempo que é muito diferente do anime rotulado como "não gostou". Nesse sentido, casos onde o conjunto de treino do usuário apresentou animes com pouca semelhança entre si ou muito semelhante com os dados rotulados como "não gostou" apresentou resultados ruins, o que é retratado no conjunto de animes rotulados pelo Michael e pelo Toshio.

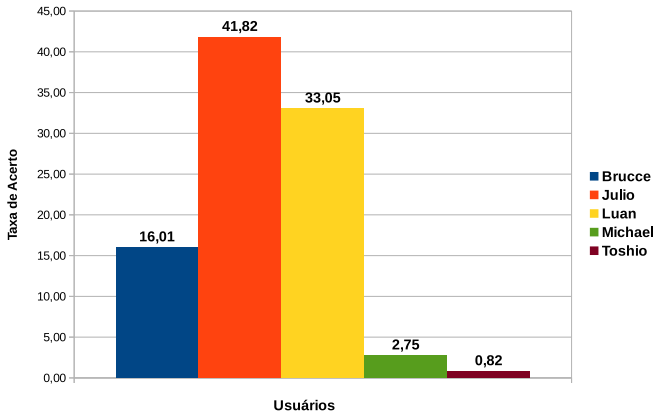

Figura 2. Resultados obtidos utilizando o primeiro critério de avaliação

O segundo critério de avaliação tem como objetivo verificar quantos dos animes que o usuário rotulou como "gostou" seriam recomendados para ele independente da posição, nesse caso um anime será recomendado se o maior valor do seu vetor F for da classe positiva. A quantidade máxima que o par de animes utilizados para gerar essa recomendação irá ganhar é a quantidade de animes que o usuário rotulou como "gostou" no conjunto de eventos rotulados. Os resultados obtidos estão representados na Figura 3,

foi observado que ao rotular animes muito semelhantes sendo um como “gostou” e outro como “não gostou”, acaba reduzindo a quantidade de acertos, como no par *Overlord III* que foi rotulado pelo usuário como “gostou” e *Munto: Toki no Kabe wo Koete* que foi rotulado como “não gostou”, uma vez que o *Overlord III* não possui descrição e está relacionado aos gêneros “Ação, Fantasia, Jogo, Magia, Sobrenatural” teve sua propagação de rótulos suprimida devido a semelhança e a representatividade do anime *Munto: Toki no Kabe wo Koete* que possui uma descrição grande e pertence aos gêneros “Ação, Fantasia, Magia e Super Poderes”, sendo muito semelhante ao *Overlord III* nos gêneros alcançando assim os mesmos animes.

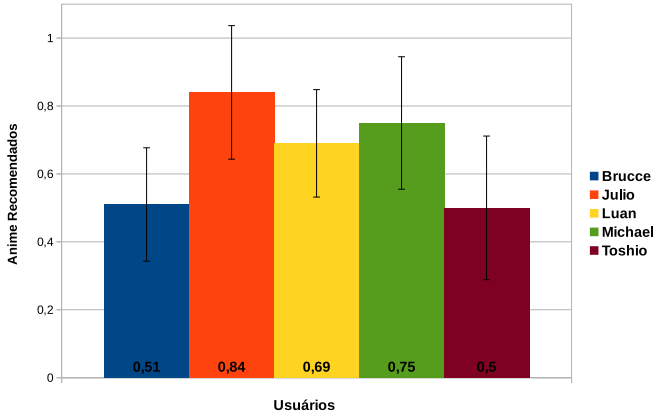

Figura 3. Resultados obtidos utilizando o segundo critério de avaliação. As barras de erros representam o desvio padrão.

5. Considerações finais e perspectivas futuras

Nesse trabalho foi proposto um sistema de recomendação de animes baseado em conteúdo, utilizando os dados textuais de suas descrições, assim como seus gêneros. Para isso, foi proposto o uso de uma técnica de aprendizado transdutivo em uma rede heterogênea, gerada a partir dos dados coletados.

De acordo com os resultados da avaliação, é possível notar que ambos os critérios utilizados são sensíveis ao conjunto de animes rotulados. Se o usuário rotular animes muito diferentes como “gostou”, ou se o animes rotulados como “não gostou” forem muito semelhantes aos rotulados como “gostou”, então o método terá poucos acertos.

Também foi observado que mesmo em um cenário extremo, com somente um par de dados rotulados (1 “gostei” e 1 “não gostei”) foi possível obter resultados que iriam satisfazer o usuário.

Trabalhos Futuros

Os resultados obtidos foram ranqueados de acordo com o valor de aptidão à classe positiva que o anime obteve, por este motivo **ranquear melhor resultados** ficará como trabalho futuro.

Devido o curto prazo para a realização de testes, os valores escolhidos para como parâmetro para o GNetMine foram obtidos através de testes manuais. Por conta disso,

é necessário investir mais tempo no trabalho de **identificar os valores ideais para os parâmetros**, tarefa que deve ficar para trabalhos futuros.

A avaliação contou somente com um usuário externo ao trabalho, devido ao tempo curto para elaboração do trabalho, sendo assim é necessário avaliar o sistema utilizando mais usuários – uma forma automática de fazer isso é através da coleta de avaliações feitas por usuários cadastrados no site *MyAnimeList*, que podem ser usadas para identificar os animes de seu gosto e assim ser usada para testes do sistema.

Referências

Blanco, R. and Lioma, C. (2012). Graph-based term weighting for information retrieval. *Information Retrieval*, 15(1):54–92.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Conference on Computational Learning Theory*, pages 92–100. ACM.

Breve, F. A., Zhao, L., Quiles, M. G., Pedrycz, W., and Liu, J. (2012). Particle competition and cooperation in networks for semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1686–1698.

Burke, R. (2007). Hybrid web recommender systems. *The Adaptive Web*, pages 337–408.

Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press.

Delalleau, O., Bengio, Y., and Le Roux, N. (2005). Efficient non-parametric function induction in semi-supervised learning. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 96–103.

Gammerman, A., Vovk, V., and Vapnik, V. (1998). Learning by transduction. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 148–155. Morgan Kaufmann Publishers Inc.

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250. ACM.

Jameson, A., Willemsen, M. C., Felfernig, A., de Gemmis, M., Lops, P., Semeraro, G., and Chen, L. (2015). Human decision making and recommender systems. *Recommender Systems Handbook*, pages 611–648.

Ji, M., Sun, Y., Danilevsky, M., Han, J., and Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Springer.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1 edition.

Motta, C. L. D., Garcia, A. C. B., Vivacqua, A. S., Santoro, F. M., and Sampaio, J. O. (2011). *Sistemas de recomendação*. Elsevier.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc.

Reategui, E. B. and Cazella, S. C. (2005). Sistemas de recomendaçao. In *XXV Congresso da Sociedade Brasileira de Computação*, pages 306–348.

Resnick, P. and Varian, H. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

Ricci, F., Rokach, L., and Shapira, B. (2015). Recommender systems: introduction and challenges. *Recommender Systems Handbook*, pages 1–34.

Rossi, R. G., de Andrade Lopes, A., and Rezende, S. O. (2016). Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. *Information Processing & Management*, 52(2):217–257.

Rossi, R. G., Lopes, A. A., and Rezende, S. O. (2014). A parameter-free label propagation algorithm using bipartite heterogeneous networks for text classification. In *Proceedings of the 29th ACM Symposium on Applied Computing*, pages 79–84. ACM.

Rossi, R. G., Rezende, S. O., and de Andrade Lopes, A. (2015). Term network approach for transductive classification. In *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 497–515.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Silva, T. C. and Zhao, L. (2012). Stochastic competitive learning in complex networks. *IEEE Neural Networks and Learning Systems*, 23(3):385–398.

Soares, M. and Viana, P. (2015). Tuning metadata for better movie content-based recommendation systems. *Multimedia Tools and Applications*, pages 7015–7036.

Sun, Y. and Han, J. (2012). *Mining Heterogeneous Information Networks: Principles and Methodologies*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers.

Wang, F. and Zhang, C. (2006). Label propagation through linear neighborhoods. In *Proceedings of the 23th International Conference on Machine Learning*, pages 985–992. ACM.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency. In *Advances in Neural Information Processing Systems NIPS*, volume 16, pages 321–328.

Zhu, X. (2005). Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919. AAAI Press.

Zhu, X. and Goldberg, A. B. (2009a). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

Zhu, X. and Goldberg, A. B. (2009b). *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers.