UNIVERSIDADE FUMEC
FACULDADE DE CIÊNCIAS EMPRESARIAIS
MESTRADO EM SISTEMAS DE INFORMAÇÃO E GESTÃO DO CONHECIMENTO

ARMSTRONG GOMES BRITO

PROPOSTA DE MODELO DE RECOMENDAÇÃO DE CONTEÚDO BASEADO EM ARQUIVOS DE LEGENDAS DE FILMES E SÉRIES

Belo Horizonte
2016

ARMSTRONG GOMES BRITO

PROPOSTA DE MODELO DE RECOMENDAÇÃO DE CONTEÚDO BASEADO EM ARQUIVOS DE LEGENDAS DE FILMES E SÉRIES

Dissertação apresentada ao curso de Mestrado em Profissional em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC, como requisito parcial para a obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento.

Área de concentração: Gestão de Sistemas de Informação e Gestão do Conhecimento.

Linha de Pesquisa: Tecnologia e Sistemas de Informação

Orientador: Prof. Dr. Luiz Cláudio Gomes Maia

Belo Horizonte
2016

Brito, Armstrong Gomes.

B862p Proposta de modelo de recomendação de conteúdo baseado em arquivos de legendas de filmes e séries / Armstrong Gomes Brito. — Belo Horizonte, 2016.
x, 102 f.: il.; 29 cm.

Dissertação (mestrado) - Universidade FUMEC – Faculdade de Ciências Empresariais.

Orientador: Luiz Cláudio Gomes Maia

1. Sistemas de informação – Teses. 2. Sistemas de recomendação 3. Sistemas de recuperação da informação. I. Orientador. II. Título.

CDU 658.403 (043)

UNIVERSIDADE
FUMEC
DE MINAS GERAIS PARA O MUNDO

Dissertação intitulada "**Proposta de modelo de recomendação de conteúdo baseado em arquivos de legendas de filmes e séries**" de autoria de Armstrong Gomes Brito, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Luiz Cláudio Gomes Maia – Universidade FUMEC
(Orientador)

Prof. Dr. Alair Dias Júnior – Universidade FUMEC
(Examinador Interno)

Profa. Dra. Gercina Ângela Borém Lima – UFMG
(Examinador Externo)

Gracielle Mendonça Rodrigues Gomes, Me. – UFMG
(Consultor *Ad Hoc*)

Prof. Dr. Fernando Silva Parreiras
Coordenador do Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC

Belo Horizonte, 25 de maio de 2016.

Rua Cobre, 200 - Cruzeiro
30310-190 - Belo Horizonte- MG
Tel. (31) 3228-3060
www.fumec.br

RESUMO

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem modelos de recuperação e recomendação cada vez mais sofisticados. O objetivo deste trabalho é propor um modelo de recomendação de conteúdo baseado em arquivos de legendas de filmes e séries. Utilizando a ferramenta *Apache Luciene* para recuperação da informação e a ferramenta OGMA para análise de textos, foi possível propor, para o modelo, três etapas distintas: uma pesquisa utilizando palavra-chave, a classificação de filmes e séries por gênero e a identificação de títulos similares. Também é apresentada uma adaptação ao modelo para identificar em cada título um sentimento denominado análise de sentimentos. Como resultado, ressaltamos que a pesquisa por palavras-chave gerou recomendações surpreendentes, já que proporcionam ao usuário liberdade de pesquisa dentro de um conteúdo específico. Já a classificação por gênero apresentou índice de 73% de acerto em comparação com os gêneros apresentados pelo site IMDb, facilitando a recomendação de conteúdo. A análise de sentimentos demonstrou recomendações com coesão, determinando títulos apropriados para cada sentimento. Por último, a identificação de títulos similares, apresentou resultados primários, trazendo apenas filmes e séries com a mesma temática, sem apresentar nenhum resultado em comum com o site IMDb. Concluiu-se que apesar da enorme dificuldade de ser assertivo na recuperação da informação, existem vantagens em se utilizar os arquivos de legendas para ajudar na composição dos sistemas de recomendação.

Palavras-chave: Recomendação de conteúdo; Recuperação da informação; Recomendação de filmes e séries; Arquivos de legenda; Classificação por gênero; Apache-Lucene; OGMA; Sistemas de recomendação.

ABSTRACT

The growing complexity of stored objects and the large volume of data requires recovery models and recommendation increasingly sophisticated. The objective of this work is to propose a content recommendation model based on movie subtitle files and series. Using Lucene tool for information retrieval, and OGMA tool for text analysis, it was possible to propose to the model, three distinct steps: a search using word-keys, the classification of films and series by genre and identification of similar securities. It is also presented an adaptation of the model to identify in each title a feeling, called sentiment analysis. As a result we note that the search for keywords generated amazing recommendations, as they provide the user freedom to search within a specific content. Already the gender breakdown showed index of 73 % accuracy compared to the genres presented by IMDb, facilitating the recommendation of content. The analysis showed feelings recommendations cohesion , determining appropriate titles for each feeling. Finally, the identification of similar titles, presented primary results, bringing only films and series with the same theme, without showing any results in common with the IMDb. It was concluded that despite the enormous difficulty being assertive in information retrieval, there are benefits to using the subtitle files to help in the composition of recommendation systems.

Keywords: Content recommendation; Information retrieval; Recommendation of movies and series; Subtitles files; Gender breakdown; Apache-Lucene; OGMA; Recommender systems.

LISTA DE SIGLAS

API	*Application Programming Interface*
CNPq	Conselho Nacional de Pesquisa
CDN	*Content Delivery Network*
DCBD	Descoberta de Conhecimento em Bancos de Dados
DVD	*Digital Versatile Disc*
EUA	Estados Unidos da América
FBI	*Federal Bureau of Investigation*
IBM	*International Business Machines*
IDE	*Integrated Development Environment*
IDF	*Inverse Document Frequency*
IMDb	*Internet Movie Database*
IOS	*Iphone Operating System*
KDD	*Knowledge Discovery in Database*
KNN	*K-Nearest Neighbors*
MPEG	*Moving Picture Expert Group*
PwC	*PricewaterhouseCoopers*
RI	Recuperação da Informação
SMART	*Sistem for the Manipulation and Retrieval of Text*
SRI	Sistema de Recuperação da Informação
TF	*Term Frequency*
VIRUS	*Video Information Retrieval Using Subtitles*
WEB	*World Wide Web*

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. Para Choo e Rocha (2003), recuperar uma informação é disponibilizá-la ao usuário, que a solicita por necessidades espontâneas e/ou induzidas, objetivando construir significado, produzir novo conhecimento e tomar decisões, sejam administrativas, sejam pessoais. Já Mooers (1951), define que a recuperação da informação trata dos aspectos intelectuais da descrição da informação e sua especificação para a busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação. A recuperação de informação apresenta a cada dia, novos desafios.

> Os volumes de dados de sistemas modernos de *software* estão crescendo cada vez mais. Empresas como o *Google* têm indexado mais de 10 bilhões de registros; DVDs (*Digital Versatile Disc*) de imagem de alta qualidade de filmes contêm *gigabytes* de dados; sistemas de dados científicos têm uma enorme coleta de dados (*terabytes*) de instrumentos científicos de alta resolução (MATTMANN et al., 2007).

Uma boa estratégia de busca normalmente assegura resultados mais relevantes do que apenas resultados satisfatórios. A insatisfação com os resultados das buscas pode estar relacionada ao pouco tempo que o usuário final dedica ao preparo de sua estratégia de busca. DiMartino e Zoe (1996) apresentaram um estudo sobre as técnicas de estratégia de busca utilizadas pelo usuário final treinado, para verificar se eles estão ou não capacitados para executar as buscas efetivamente: concluíram que 55% dos usuários finais ficam insatisfeitos com suas próprias buscas. A busca pela informação deve ser de fácil acesso ao usuário para que ele gaste o menor tempo possível e encontre exatamente o que ele está procurando. As ferramentas devem ajudar o usuário, devem ser intuitivas e não confundi-los.

A experiência dos usuários na utilização dos mecanismos de busca nem sempre é positiva. Fernandes et al. (2013) cita que quatro em cada dez usuários na pesquisa relatam não só resultados contraditórios que impossibilitavam descobrir qual a informação correta, como também um excesso de informações nos resultados que os deixavam atônitos (sobrecarregados). Além disso, um em cada três apontou ausência de informações importantes ou essenciais nos resultados apresentados.

Além da estratégia de busca, um processo que visa ajudar e dar destaque ao que possa ser informação relevante para o usuário é a indexação. Primeiro é armazenado o que se deseja para posteriormente ser recuperado. Novas técnicas têm sido desenvolvidas, através da indexação, para resolver de forma mais efetiva o principal problema dos sistemas de recuperação: a relevância para o usuário. No estado atual da pesquisa, a indexação automática ainda apresenta limitações. É necessário reconhecer os diferentes significados e relações entre os conceitos, e desenvolver algoritmos que considerem a semântica e a sintaxe do conteúdo dos documentos.

Um dos mercados que necessitam utilizar as ferramentas de recuperação de informação é o mercado de *streaming*, utilizado hoje por grandes empresas na Internet. A tecnologia de streaming é uma forma de transmissão instantânea de dados de áudio e vídeo através da Internet. Por meio desse serviço, é possível assistir a filmes, séries ou mesmo escutar músicas sem a necessidade de realizar o *download*, o que torna mais rápido o acesso aos conteúdos online.

Uma das empresas que realizam a exploração da tecnologia de *streaming* é o *Netflix*. O *Netflix* é um site na internet para assistir filmes, séries e documentários. Ele utiliza o conceito *on-demand*, que pode ser entendido, basicamente, em ter a disponibilidade de assistir aos programas que o usuário deseja. Isto é possível devido à tecnologia a cabo e um servidor local, que faz o streaming das imagens. No modelo de negócio utilizado é cobrado uma assinatura mensal para que sejam disponibilizados filmes e séries que podem ser assistidos em diversos dispositivos (computadores, celulares e *tablets*) a qualquer hora. Os serviços de *streaming on-demand* possibilitam que o usuário esteja no controle do que vai assistir, quando e onde.

Segundo Adhikari et al. (2012) o *Netflix* representa 29,7% do tráfico de Internet nos Estados Unidos. No Brasil, o consumo *on-demand* cresce 50% ao ano, segundo dados do Instituto de Pesquisa Dataxis (DATAXIS, 2015). Tudo isso requer uma infraestrutura adequada e pensada de forma inteligente. São várias as técnicas utilizadas para que o usuário assista com qualidade o título que deseja sem congestionar as redes das operadoras.

Adhikari et al. (2012) cita que a arquitetura do *Netflix* é composta por múltiplas redes de entrega de conteúdo (CDNs - *Content Delivery Network*) para tudo seja entregue aos usuários finais com qualidade e performance. Adhikari et al. (2012) fazem uma investigação de como cada CDN é selecionado, dependendo da localização geográfica do usuário, além de analisarem e proporem estratégias para obter uma melhor performance. As CDNs evitam que todos os usuários finais tenham que ir até um único servidor em outro país. Isso demonstra a

dificuldade enorme de recuperação de conteúdo para cada título do acervo. Sistemas de recomendações devem ser idealizados já partindo deste pré-suposto. A utilização de um arquivo leve, como o arquivo texto das legendas, atende a esta expectativa, porque é um arquivo pequeno, de poucos bytes, que pode ser recuperado rapidamente mesmo que o usuário esteja distante do servidor.

Outras empresas oferecem serviços parecidos e possibilitam também o *streaming* de músicas, como o *Spotify* e recentemente o *Google* e a *Apple*. São várias as possibilidades também no mercado das músicas. Encontrar suas músicas favoritas, ver as *playlists* de seus amigos ou mesmo encontrar aquela música perdida no tempo são umas das opções. Utilizar a letra da música para criar sistemas de recomendação de conteúdo pode ser promissor. Este trabalho apesar de ser aplicado na área de filmes e séries poderá, se adaptado, encaixar provavelmente para a área musical.

> Um ano faz uma diferença e tanto! No final de maio de 2014, atingimos 10 milhões de assinantes pagos e 40 milhões de usuários ativos. Hoje nós alcançamos mais de 20 milhões de assinantes e mais de 75 milhões de usuários ativos. Foram 10 milhões de assinantes durante nossos primeiros cinco anos e meio – e outros 10 milhões de assinantes em apenas um ano! Isso dá uma média de um novo assinante a cada 3 segundos no último ano. Wow! Mais pessoas usando o *Spotify* significa mais pagamento para os criadores das músicas que você ama. Enquanto nós crescemos, o montante de *royalties* que pagamos aos artistas, compositores e detentores de direito continua a subir, cada vez mais rápido. Nós já pagamos mais de USS3 bilhões em *royalties*, incluindo mais de USS300 milhões nos primeiros meses de 2015 (SPOTIFY, 2015).

Os fatos demonstram que apesar do mercado em expansão, há ainda um longo caminho para percorrer, inúmeros usuários para serem conquistados, a difícil superação dos problemas de infraestrutura, entre outros obstáculos, mas os serviços de *streaming* já são destaques no mercado brasileiro. Um problema frequente que estes consumidores e de outros serviços enfrentam é a assertividade de escolha do conteúdo. O melhor filme ou série para um usuário nem sempre é aquele aclamado pela crítica, o mais assistido ou inda o ganhador de mais estatuetas do *Oscar*. O filme preferido do usuário pode ser simplesmente aquele que mais lhe emocionou, comoveu ou se sentiu contagiado pela história. Os elementos daquela série ou filme podem fazer usuários diferentes sentirem diferentes sensações, apesar de ser o mesmo título. A pesquisa por um filme torna-se desta forma única e muito subjetiva. É um obstáculo que as grandes empresas terão que superar evoluindo e inovando cada vez mais suas ferramentas.

A busca padrão por nome do filme, diretor, ano de lançamento, os filmes mais assistidos, entre outros, muitas vezes não atende completamente ao usuário. A busca por estes

metadados é importante, mas não atende todos os requisitos de um usuário mais exigente. Outras alternativas são realizadas com a ajuda das redes sociais, como a indicação de filmes que seus amigos já assistiram. Apenas recentemente, por exemplo, o *Netflix* atualizou sua página de buscas conforme comentário em seu blog oficial:

> No *Netflix*, estamos sempre pensando em novas formas de melhorar a experiência do usuário. Hoje lançamos uma nova experiência de busca no site do *Netflix*, que complementa as sugestões de filmes e séries do sistema de recomendações. Com a nova ferramenta de busca instantânea, você terá acesso mais rápido e fácil aos seus filmes, séries de TV e atores favoritos. Uma lista de atores, diretores e criadores correspondentes à busca é exibida no lado esquerdo. Ao clicar em um dos resultados, você verá uma nova galeria com todos os títulos relacionados à pessoa selecionada. Em cada caso, os usuários verão recomendações de acordo com o que muitos assinantes decidiram assistir depois de fazer aquela mesma busca (NETFLIX, 2015).

A disponibilização para o usuário de uma ferramenta que permita a interação com o conteúdo poderá trazer inúmeros resultados satisfatórios, criando meios de descobrir novos títulos. O conteúdo é a fonte mais fidedigna do título, é ele que o autor vai ver, escutar, arrepiar, rir, enfim, sentir as mais variadas emoções. Com o tempo cada vez mais escasso as pessoas devem ter assertividade nas suas escolhas para que os resultados das pesquisas tragam realmente o que desejam. Quem nunca começou a assistir um filme e já no começo do mesmo já estava decepcionado? A preocupação mais com o conteúdo do que com os metadados é uma das virtudes que os sistemas de recuperação da informação devem se preocupar.

> Agora, ao selecionar um título, o aplicativo passará a reproduzir o título enquanto exibe as informações mais relevantes para ajudar você a confirmar sua escolha. A inovação dá maior destaque ao conteúdo em si e proporciona uma experiência mais imersiva e cinemática, digna dos recursos e expectativas das TVs atuais. O vídeo é um meio rico para se contar histórias, e descobrimos que a reprodução de títulos durante a seleção de conteúdo facilita ainda mais a busca por algo interessante para assistir (NETFLIX, 2015).

Com a quantidade de informações e com a disponibilidade facilitada destas informações pelo acesso à Internet, as pessoas se deparam com uma diversidade muito grande de opções. Muitas vezes um usuário possui pouca ou quase nenhuma experiência pessoal para realizar escolhas dentre as várias alternativas que lhe são apresentadas. Pode ser difícil tomar uma decisão. A questão relevante neste momento refere-se a como proceder nestes casos? Para diminuir as dúvidas e necessidades que temos frente à escolha entre inúmeras alternativas, geralmente confiamos nas recomendações que são passadas por outras pessoas ou através de

textos de recomendação, opiniões de revisores de filmes e livros, sites da Internet, impressos de jornais, dentre outros.

Um exemplo de recomendação existente hoje na Internet e utilizado por milhares de usuários é o site denominado *Internet Movie Database* (IMDb) disponível no endereço *http://www.imdb.com/*. O site não disponibiliza filmes ou séries para assistir via streaming, como no *Netflix*, muito menos permite realizar o *download* do título, mas é uma excelente fonte de informação. Ele possui um dos maiores e mais completo banco de dados sobre cinema e TV do mundo. Ele existe tanto em formato de site, como de aplicativo (grátis) para *Android* ou IOS (*Iphone Operating System*).

Dentro do site IMDb o usuário pode navegar e achar um enorme volume de informações a cada filme, diretor ou ator ou pesquisado. Uma busca pode começar a partir de um simples trailer, para em seguida as informações serem detalhadas de diversas formas, como: o gênero do filme, o ano, o tempo de duração, a média da nota dada pelos usuários do site, nome do diretor, o nome dos personagens e seus respectivos atores, um resumo sobre o filme, sites oficiais e curiosidades, ou seja, as alternativas são enormes dentro do site.

> IMDb é a fonte mais popular do mundo para pesquisar sobre filmes, TV e celebridades. O site IMDb (*http://www.imdb.com/*) é número 1 do mundo, nestes quesitos, com uma audiência de mais de 250 milhões de visitantes únicos mensais. IMDb oferece um banco de dados de mais de 185 milhões de itens de dados, incluindo mais de 3 milhões de filmes, programas de TV e de entretenimento e mais de 6 milhões de elenco e membros da tripulação. Os consumidores confiam na informação que o IMDb fornece - incluindo horários de exibição de filmes locais, bilheteria, críticas e comentários, recomendações personalizadas, galerias de fotos, notícias de entretenimento, citações, curiosidades, dados de bilheteria e uma página para decidir quando e o que assistir, além de onde vê-lo (IMDB, 2015).

No site IMDb (2015), eles relatam que provêm aplicativos de entretenimento como os populares *Movies* e TV para *iPhone*, *iPad*, *Kindle Fire*, telefones *Android*, *tablets Android* e tem também seu site otimizado para acesso por *smartphones*. Ainda segundo o IMDb (2015), houve mais de 115 milhões de downloads de aplicativos móveis da IMDb em todo o mundo. Este volume de *downloads* demonstra o quanto as pessoas estão interessadas no conteúdo de filmes e séries. Além disso, o IMDb realiza o programa denominado *What to Watch* (*http://www.imdb.com/whattowatch*), para ajudar a fãs descobrir e mergulhar profundamente nos seus filmes e séries preferidos.

Ao realizar uma busca avançada o site oferece inúmeros filtros, oferecendo aos usuários uma enorme gama de possibilidades para que seja possível explorar todo seu conteúdo.

Nas figuras 1 e 2, são apresentados algumas das possibilidades de pesquisa, podendo o filtro ser realizado: pelo título, pelo tipo do título, pela data de lançamento, pela nota dos usuários, pelo número de votos, pelo gênero, pelo país, pela língua, entre outros.

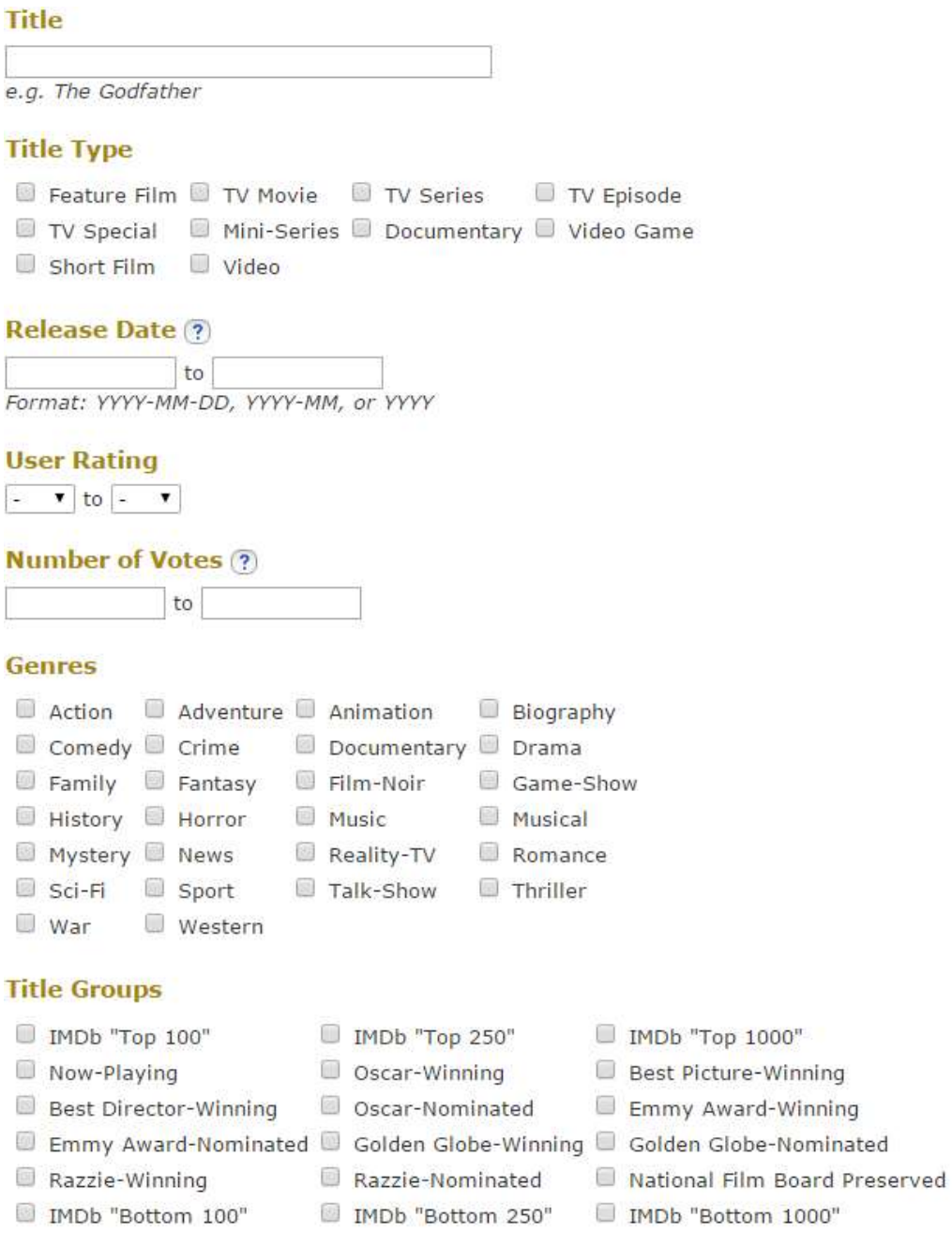


Figura 1: Busca avançada IMDb
Fonte: IMDb (2015)

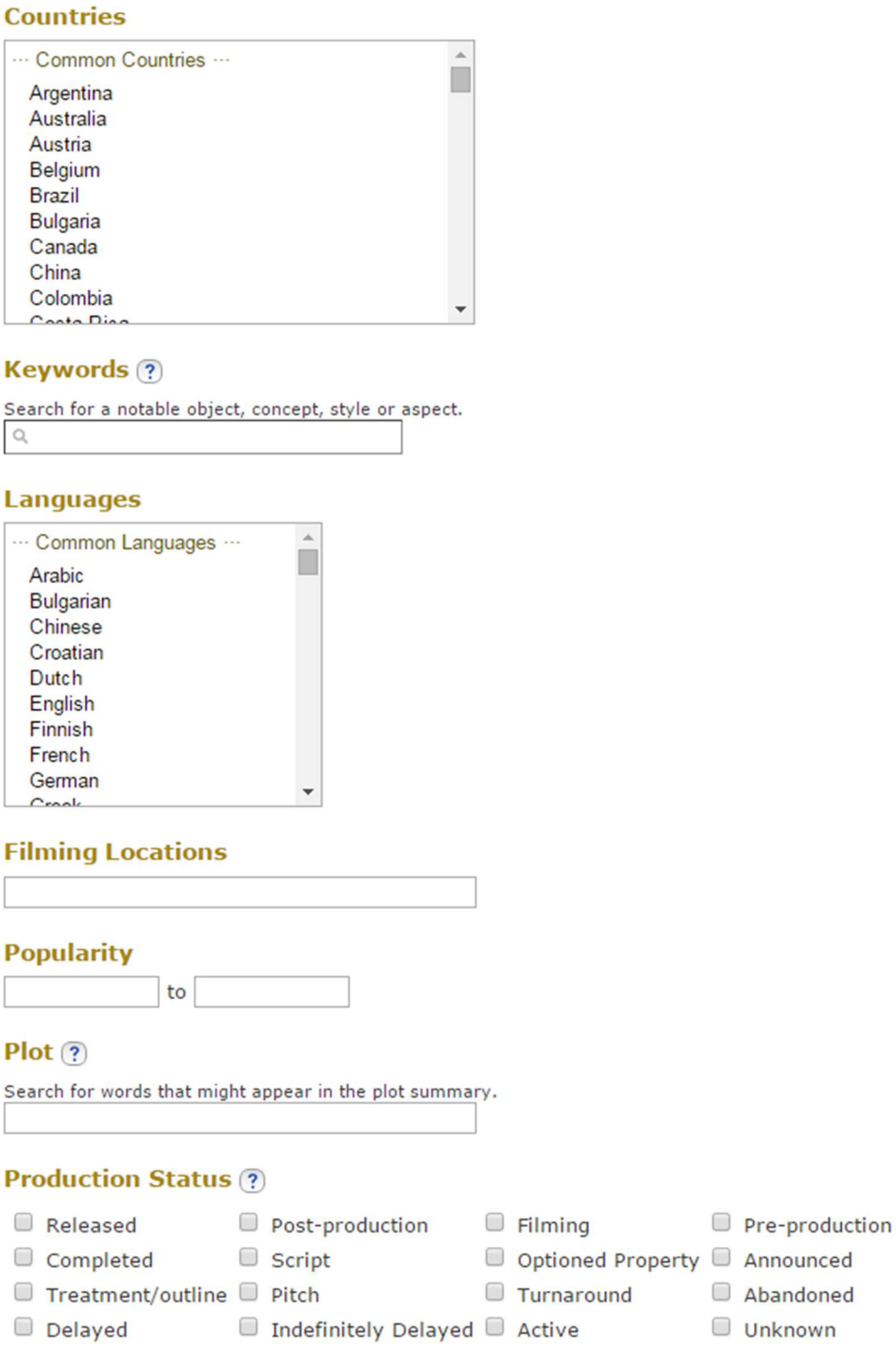


Figura 2: Outros filtros de busca – IMDb
Fonte: IMDb (2015)

Outra funcionalidade do *site* é a possibilidade de criação de um perfil. Ao criar um perfil, é possível classificar através de notas de 0 a 10 todos os filmes e seriados que já assistiu. Com esta classificação o IMDb consegue expressamente recomendar títulos personalizados para o seu perfil. Ele utiliza destas classificações para comparar seus dados com as avaliações feitas por outros usuários. Desta forma, serão encontradas pessoas com o gosto semelhante ao

usuário, identificando filmes que você ainda não assistiu. Para cada recomendação, você pode ver uma lista dos filmes ou programas de TV em que a recomendação foi baseada.

Apesar das mais variadas vantagens que o site apresenta, ele deve ser atualizado constantemente e depende de uma equipe enorme para que o mesmo continue sendo relevante. Além disso, o sistema de recomendação utilizado depende de o usuário completar seu perfil e avaliar os filmes e séries que já assistiu para que o sistema de recomendação de títulos seja eficiente.

Diante deste cenário, este trabalho propõe um modelo de recomendação de conteúdo que facilite a busca por conteúdo em filmes e séries, sem que dependa de inúmeras interações realizadas anteriormente pelo usuário. Este modelo utilizará como base de dados os arquivos de legendas de cada título (filmes ou séries). Todos estes itens ficam "pelo caminho" nos sistemas tradicionais e estruturados de buscas realizadas através dos metadados. Com o aumento do acervo e variedade de lançamentos que ocorrem, torna-se cada vez mais necessário a organização desses itens para uma recuperação eficiente. A proposta deste trabalho é justamente ampliar a possibilidade de busca e garantir que o usuário tenha em mãos uma ferramenta para encontrar justamente o que vem procurando naquele momento.

1.1 Problema

Devem existir alternativas de pesquisa além do padrão de pesquisa realizada hoje pelo nome do filme, diretor, ano de lançamento, os filmes mais assistidos, entre outros. Hallinan e Striphas (2016) citam que as informações pessoais como o sexo, idade, raça e outras classificações gerais não são informações relevantes para recomendações.

Além disso, sistemas de recomendação dependem do perfil do usuário e da classificação que ele executa a cada filme assistido. As redes sociais são outra tentativa utilizada para recomendação, mas quem garante que os filmes que seus amigos já assistiram, podem ser recomendados para você? Recomendações de conteúdo fornecidas a partir do conteúdo dos títulos são mais complexos, logo mais modelos devem ser propostos e estudados.

Desta forma, a questão problema que irá nortear esta dissertação será "Quais as possibilidades e/ou alternativas de recomendação de conteúdo utilizando como base de dados os arquivos de legendas de filmes e séries?".

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo geral do trabalho é propor um modelo de recomendação de conteúdo baseado nos arquivos de legendas de filmes e séries.

1.2.2 Objetivos específicos

- Criar um mecanismo de busca, a partir de palavras-chave, que resulte em uma relação de títulos classificados pela sua relevância;
- Classificar os títulos por gênero através da análise de conteúdo dos arquivos de legenda;
- Recomendar títulos similares, a partir de um único título pré-definido.

1.3 Justificativa

A indústria cinematográfica, os provedores de filmes e séries (como o Netflix e Net-Now) entre outras empresas deste ramo, necessitam que uma maior gama de filmes sejam encontrados e utilizados, já os usuários desejam descobrir inúmeros filmes interessantes e raros que a maioria ainda não assistiu, ou apenas mais alternativas de pesquisa para encontrar o que deseja dentro de um grande acervo. O projeto propõe um modelo de recuperação de informação baseada em conteúdo. Segundo Foskett (1972 apud SOUZA et al., 2006); a recuperação de informações traz dificuldades intrínsecas ao conceito de “informação”, como a dificuldade da determinação da real necessidade do usuário e do eu melhor atendimento com os documentos

que fazem parte do acervo do sistema. Como garantir que os indivíduos recebam informações relevantes e atualizadas em qualquer busca realizada?

Para Souto (2012 apud PINHEIRO; LOUREIRO, 1995), Saracevic é o autor que melhor identifica o sentido de informação dentro da área de Ciência da Informação, distinguindo informação e informação relevante, relacionando esta última a mecanismos de comunicação seletiva e à orientação aos usuários de sistemas de recuperação da informação. A informação relevante deveria ser a premissa para os sistemas de recuperação de informação, entretanto, a dificuldade é identificar quando uma mesma informação é relevante para um grupo de usuários e não para outro. Barreiro (1978) já se preocupava com a grande quantidade de informação disponível o que tornava quase impossível sua identificação, alertando sobre a necessidade de um mecanismo que filtrasse as informações de modo que o usuário tivesse conhecimento apenas das informações potencialmente relevantes. Para Souto (2003), à medida que as técnicas de tratamento da informação têm progredido, surgiram novos serviços mais adequados às necessidades dos usuários.

A Internet vem evoluindo a maneira que interage com seus usuários. Hoje a força colaborativa vem ganhando cada vez mais espaço. Velsen e Melenhorst (2009) discorrem um pouco sobre este tema.

> A *web* (*World Wide Web*) 2.0 trouxe uma tendência na Internet no qual um crescente número de sites oferece aos usuários a possibilidade de contribuir ativamente com conteúdo. Na *web* 1.0, o conteúdo era gerado exclusivamente pelo fornecedor, já na *web* 2.0, fornecedores ofertam aos usuários da web a possibilidade de criar seu próprio site/conteúdo. Exemplos deste conceito são *YouTube* e a *Wikipedia*, sites que, respectivamente, permitem aos usuários criar seus próprios banco de dados com filmes ou sua própria enciclopédia *on-line* (VELSEN; MELENHORST, 2009).

São necessárias novas ideias e técnicas para averiguar quanto cada uma pode ser relevante para ajudar no tratamento da informação. Através destas ideias novos conceitos tendem a ser criados para melhorar os sistemas de recuperação da informação. Souza et al. (2006) cita que outras metodologias similares implantadas em SRIs (Sistemas de Recuperação da Informação) permitem a busca de expressões regulares, ou mesmo analisam a proximidade da ocorrência de alguns termos, expandindo o conceito de palavra-chave para frases ou outras hierarquias lexicais. Algumas estratégias de recuperação da informação são apresentadas no mapa conceitual da figura 3. Esta figura demonstra os caminhos possíveis para estruturar e melhorar os sistemas de recuperação da informação.

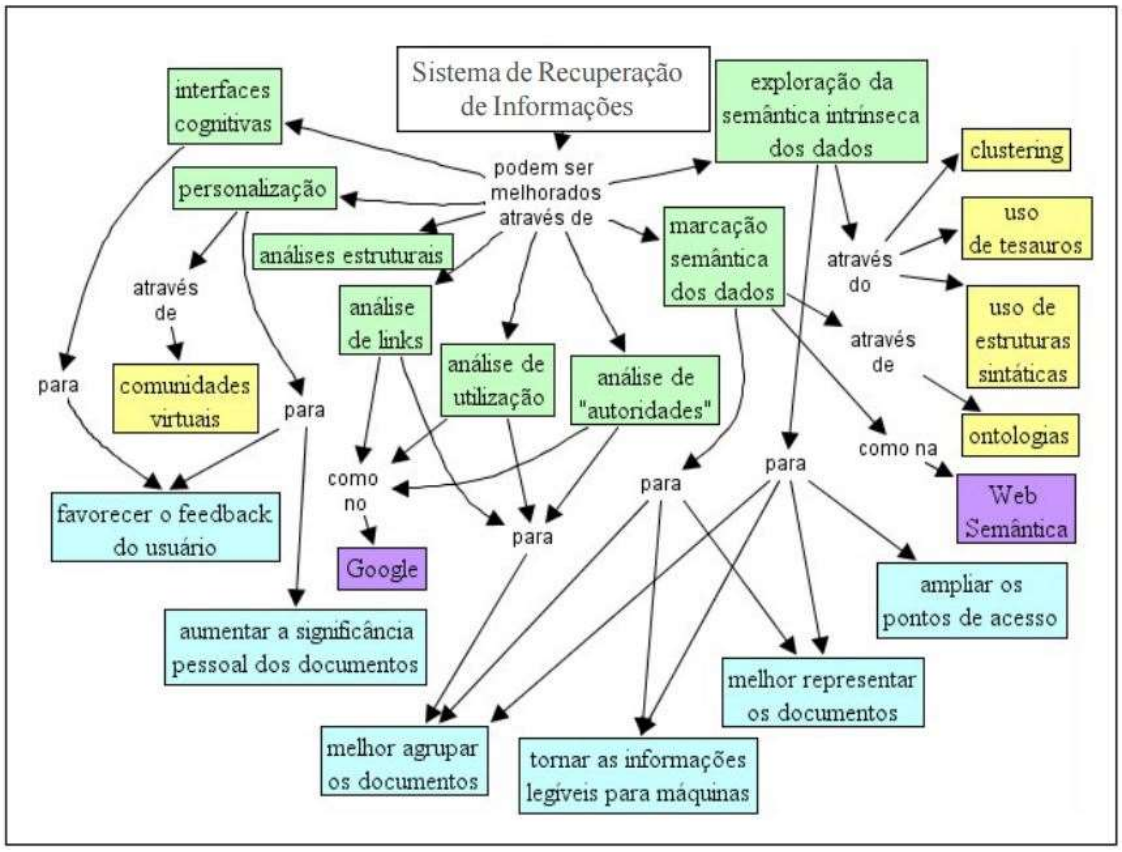


Figura 3: Sistema de recuperação de informações
Fonte: Souza et al (2006)

Baeza-Yates e Ribeiro-Neto (1999) citam que um dos problemas centrais da recuperação de informações em SRIs é a predição de quais são os documentos relevantes e quais devem ser descartados, e essa tarefa de "escolha", em sistemas automatizados, é executada por algum tipo de algoritmo que, baseado em heurísticas previamente definidas, decide quais são os documentos relevantes a serem recuperados e os ordena a partir dos critérios estabelecidos. Souza et al. (2006) citam ainda que na indexação automática, existem dezenas de estratégias para a correta ponderação do valor dos documentos, de acordo a necessidade de informação. A grande dificuldade é encontrar o meio termo, do que pode ser realizado automaticamente pelos softwares e o que deve ser realizado manualmente pelas pessoas. Achar este ponto de equilíbrio é um ponto de discussão dentro da academia. Souza et al. (2006) enfatizam que quando a indexação é realizada manualmente – ou melhor expressando, intelectualmente – por seres humanos, cabe a estes descobrir conceitos que sirvam de termos-índices para serem vasculhados durante as consultas (queries) de usuários.

Na et al. (2011) citam que muitos estudos de classificação automática de texto têm centrado em classificação de documentos por assunto ou tópico (por exemplo, educação versus

entretenimento), ou classificação de gênero. Este trabalho abrange além da classificação por gênero, a pesquisa e a similaridade de documentos baseado em conteúdo.

> O surgimento e a ampla utilização de bibliotecas digitais impõe muitos novos desafios. É cada vez mais importante inovar a forma de organizar as informações e apoiar o usuário a encontrar documentos relevantes. As bibliotecas digitais oferecem boas oportunidades como ambientes de teste para explorar estes problemas com vários tipos de documentos digitais. Portanto, os pesquisadores estão desenvolvendo novas formas de organizar, navegar e buscar documentos digitais, não só pelos campos de metadados padrão, mas também por aspectos dos documentos através de conteúdo e apoiado em um sistema automático de análise (NA et al., 2011).

O conceito de disseminação seletiva de informações (SDI) é citado por Campos (2012 apud LUHN, 1958) ao reportar o desenvolvimento de um sistema de inteligência de negócios na IBM (International Business Machines) que pretendia criar perfis de unidades ou indivíduos na organização que determinassem o recebimento de metadados gerados automaticamente sobre documentos que provavelmente lhes despertariam o interesse. Podemos observar claramente o objetivo do mercado em entregar a informação correta para cada grupo de usuários. O objetivo do sistema era prover informação adequada para fornecer suporte às atividades executadas por indivíduos, grupos ou unidades maiores, os chamados pontos de ação, criando meios para disseminar informação para cada um dos pontos de ação de acordo com suas requisições e desejos (CAMPOS, 2012 apud LUHN, 1958).

Segundo Schafer, Konstan e Riedi (1999), o aumento da possibilidade de disponibilização de conteúdo (produtos ou informação) através de sistemas web faz com que os sistemas web apresentem mais opções de escolha para o usuário antes de ele estar apto a encontrar a opção que vai ao encontro de sua necessidade. Uma das soluções para resolver este problema de sobrecarga de informação é a utilização de Sistemas de Recomendação.

Segundo Reis e Pereira (2010), alguns pesquisadores e colaboradores da rede mundial de computadores estão tentando facilitar o acesso a essa grande quantidade de informação que está disponível e descentralizada na web. Estudos e técnicas como a Inteligência Artificial, a web Semântica, e a Indexação de Arquivos junto com Algoritmos de Busca, estão sendo desenvolvidos com esse objetivo (REIS, 2011).

> A necessidade de equilíbrio na adoção da linguagem natural nos novos contextos. É instrutivo - especialmente tendo em conta o interesse recente e atividade voltada para organizar a informação digital – compreender certas características dos sistemas tradicionais bibliográficos. Dois aspectos, em particular, devem ser considerados. Um refere-se ao fato desses sistemas fornecerem soluções para os problemas que obstruem o acesso eficiente à informação. Ainda hoje alguns problemas são causados pelo

> acesso as novas tecnologias, outros - decorrem da variedade de informações, das múltiplas facetas dos usuários e das anomalias que caracterizam a linguagem adotada na recuperação da informação (MOURA, 2009).

Apesar das mudanças ocorridas nos processos de produção, tratamento e disseminação de informação, alguns problemas enfrentados pelos sistemas tradicionais de recuperação da informação continuam presentes nas ferramentas de busca atuais e ganham maior amplitude e complexidade (BRASCHER, 2002). Segundo Brascher (2002 apud CHEN, 1999), isto se deve a diferentes fatores: variações nas estruturas e formatos de bases de dados, diferentes formas de documentos disponibilizados (texto, áudio e vídeo) e abundância de conteúdos multilíngues nas aplicações da web.

O mercado de streaming é um dos mercados que utiliza os sistemas de recuperação da informação. Os consumidores de filmes e séries estão cada vez mais exigentes, o que justifica a necessidade de se ter bons mecanismos de recomendação para não deixar o usuário assistir algo que, em princípio, não será interessante. Além disso, o usuário pode ter a possibilidade de extrair consultas mais detalhadas sobre as características de um filme ou de uma série. Além do consumidor de informação, também os profissionais da indústria de cinema podem ser beneficiados com sistemas que lhes permitem acessar a cenas com características idênticas, para gerar mais ideias e possibilidades, efetuando corretamente uma comparação.

> Um estudo realizado *PricewaterhouseCoopers* (PwC) revelou que o mercado de DVDs e *Blu-ray* está declinando rapidamente, sendo preenchido por serviços de streaming, como *Netflix*, que também abocanharão as receitas de bilheterias dos cinemas nos próximos anos. De acordo com o levantamento, a receita com esse tipo de serviço (*streaming* e *download* de filmes) vai ultrapassar a mídia física em 2016, cujo mercado vai cair de US$ 12,2 bilhões este ano para US$ 8,7 bilhões em 2018. A pesquisa também prevê que, em 2017, *streaming* e *download* de filmes deixarão para trás o cinema como o maior contribuinte para a receita total de filmes nos EUA (Estados Unidos da América), chegando a US$17 bilhões em 2017, contra os atuais US$ 8,5 bilhões (THEGUARDIAN, 2015).

Segundo Souza e Alvarenga (2004), para a representação adequada de documentos, é necessário criar sistemas de indexação eficazes, de forma que a recuperação das informações neles contidas, de acordo com as necessidades dos usuários, seja a mais significativa possível. Souza e Alvarenga (2004) reafirmam que a determinação do processo de indexação é viável no momento em que os sistemas são projetados e deve funcionar continuamente, à medida que novas informações são adicionadas ao sistema.

> Em 6 de Outubro de 2006, o *Netflix*, lançou o desafio *Netflix Prize*. Tratou-se de um concurso que ofereceu 1 milhão de doláres para a primeira pessoa ou equipe que desenvolve-se um sistema capaz de predizer recomendação de filmes com pelo menos 10% que seu sistema atual. A competição atraiu mais de 50.000 participantes de 186 países, que organizaram-se em cerca de 40.000 equipes, com mais de 9000 mensagens no fórum de discussão oficial *Netflix Prize* contendo 10.000 usuários registrados, relatando os progressos e os problemas relacionados ao desafio (HALLINAN; STRIPHAS, 2016).

Cardoso (2004) conclui que a crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. Diante deste quadro, recuperação de informação apresenta a cada dia, novos desafios e se configuram como uma área de significância maior.

O mercado deve direcionar suas estratégias muito além das buscas por padrões. O usuário deve ter novas opções de escolhas e sistemas de recomendação de conteúdo eficazes. Novos filmes e séries podem ser criados de acordo com que os usuários procuram e os sistemas de recomendação podem ajudar neste caminho. O usuário não pode ser induzido a preencher inúmeros formulários, ter um cadastro e perfil completo, além da necessidade de realizar avaliações para que encontre o que esteja procurando. Diante do cenário exposto, a contribuição deste trabalho será propor um modelo que será desmembrado em três etapas: pesquisa, classificação e comparação utilizando arquivos de legendas de filmes e séries, isto tudo baseado em conteúdo. A busca por padrões e pelos metadados é realizada por inúmeros e poderosíssimos softwares, mas algumas pesquisas simples e práticas, como as indicadas neste trabalho, devem ser uma alternativa para complementar as lacunas destes sistemas.

1.4 Adequação à linha de pesquisa

A linha de pesquisa Tecnologia e Sistemas de Informação, do curso de Mestrado Profissional em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC, compreende estudos sobre conceitos e processos de desenvolvimento de tecnologias e sistemas de informação integrados com banco de dados e dotados de recursos gráficos e usabilidade avançada, de acordo com os preceitos de gestão de projetos e qualidade de *software*. Trata também dos impactos dos sistemas baseados na Internet e das novas tecnologias no comportamento do consumidor e na gestão logística.

A linha de pesquisa Tecnologia e Sistemas de Informação é a aplicação do conhecimento técnico/científico para fins de disseminação e recuperação de informações por meios computacionais. A pesquisa pretende propor um modelo de recomendação de conteúdo através de um sistema de recuperação da informação, que será utilizado para pesquisa, categorização e similaridade de conteúdo, portanto, pode-se afirmar que esta dissertação enquadra-se na área de concentração denominada de Sistemas de Informação, sob a linha de pesquisa Tecnologia e Sistemas de Informação, mais estritamente na trilha de Sistemas de Recuperação da Informação. Foram pesquisados vários conceitos referentes à recuperação da Informação, além da demonstração de um diagrama que descreve o processo de recuperação da informação. Além disso, foi dada ênfase aos sistemas de recomendação de conteúdo, seus conceitos, utilidades e características.

A trilha de Sistemas de Recuperação da Informação trata do estudo de modelos de Recuperação da Informação com foco na implementação em projetos de mineração de dados. Como o termo indica, mineração de dados refere-se à mineração ou descoberta de novas informações em termos de padrões ou regras com base em grandes quantidades de dados. Outros estudos envolvem métricas e métodos aplicáveis em redes sociais e processamento de linguagem natural. Esta temática possibilita aplicações práticas como a adaptação ou construção de ferramentas de mineração de texto existentes para usa na recuperação da informação. As aplicações de modelo de recuperação da informação são diversas, mas pode-se destacar principalmente a aplicação de análise dos dados para entender o consumidor e ajudá-lo a encontrar o que realmente deseja.

2 REFERENCIAL TEÓRICO

Este capítulo será dividido em três seções: recuperação da informação, sistemas de recuperação da informação e trabalhos relacionados. Para a seção de recuperação da informação são apresentados modelos clássicos (divididos em quantitativos e dinâmicos), e um panorama da organização da informação na web. Em seguida serão tratados os sistemas de recuperação da informação onde são apresentados os conceitos de indexação e busca. Posteriormente são discutidos os sistemas de recomendação e o conceito da análise de sentimentos. Por último será apresentada a seção dos trabalhos relacionados.

2.1 Recuperação da informação

Antes de conceituar o termo recuperação da informação é necessário conceituar o termo informação. Coadic e GOMES (1996) afirmam que a informação é um conhecimento inscrito (gravado) sob a forma escrita (impressa ou numérica), oral ou audiovisual. Eles ainda afirmam que com o advento da eletrônica, informática e do desenvolvimento da comunicação à distância, ocorreu a explosão da informação, surgindo verdadeiros supermercados da informação.

Oleto et al. (2006) ressaltam que não adianta muita informação sem qualidade. Mas o que seria qualidade da informação? Oleto et al. (2006) afirmam que a percepção da qualidade não é nítida por parte do usuário da informação, ficando mais aproximada do conhecimento popular em vez do conhecimento científico. Eles ressaltam que isso se deve provavelmente pela própria falta de conceitos claros que sustentem interpretações inequívocas da qualidade da informação (se isto for possível). Oleto et al. (2006) terminam dizendo que a tal busca, a da perfeita e precisa definição dos atributos que qualificam a informação, permitindo a inquestionável percepção do usuário, ainda não aponta para paradigmas a que se possam recorrer com segurança.

É evidente a dificuldade das pessoas em manter o foco, e trabalhar, apenas, em cima do que é necessário, em busca da qualidade da informação. Sant'ana (2008) afirma que com a adoção maciça das tecnologias de informação e comunicação, o volume de informações armazenadas e disponíveis para acesso vem crescendo de forma exponencial para que essa grande quantia de informações seja transmitida ao usuário da melhor forma, são necessários processos de recuperação cada vez mais eficientes.

A recuperação de informação (RI) pode ser considera a vertente tecnológica da Ciência da Informação e resultado da relação desta com a Ciência da Computação (SARACEVIC, 1999). De acordo com Baeza-Yates e Ribeiro-Neto (1999), a RI (Recuperação da Informação) envolve desde a representação, passando pela armazenagem, organização e chegando ao acesso aos itens da informação, promovendo, assim, facilidades de acesso do usuário à informação de interesse.

2.1.1 Modelos clássicos

Existem na literatura os modelos clássicos de recuperação da informação. Esses modelos podem ser divididos em modelos quantitativos e dinâmicos. Segundo Ferneda (2003) os modelos quantitativos além de impositivos e unilaterais, não preveem qualquer tipo de intervenção do usuário na representação dos documentos.

2.1.1.1 Modelos quantitativos

Os principais modelos quantitativos e suas características são apresentados a seguir:

• Modelo booleano: este modelo é o mais presente nos sistemas de recuperação da informação, apesar sua dificuldade em ordenar os resultados. É baseado na teoria dos conjuntos e na Álgebra de Boole (KURAMOTO, 2002).

> Um documento booleano é representado por um conjunto de termos de indexação que podem ser definidos de forma intelectual (manual) por profissionais especializados ou automaticamente, através da utilização de algum tipo de algoritmo computacional. As buscas são formuladas através de uma expressão booleana composta por termos ligados através dos operadores lógicos *AND*, *OR* e *NOT* (E, OU e NÃO) 1, e apresentam como resultado os documentos cuja representação satisfazem às restrições lógicas da expressão de busca (FERNEDA, 2003).

• Modelo vetorial: também conhecido como Modelo Espaço Vetorial, o modelo Vetorial baseia-se na comparação parcial entre a representação dos documentos e a da consulta do usuário (KURAMOTO, 2002). Uma das características importantes deste modelo é o cálculo de similaridade entre documentos.

> O modelo vetorial propõe um ambiente no qual é possível obter documentos que respondem parcialmente a uma expressão de busca. Isto é feito através da associação de pesos tanto aos termos de indexação como aos termos da expressão de busca. Esses pesos são utilizados para calcular o grau de similaridade entre a expressão de busca formulada pelo usuário e cada um dos documentos do corpus. Como resultado, obtém-se um conjunto de documentos ordenado pelo grau de similaridade de cada documento em relação à expressão de busca. (FERNEDA, 2003)

Muitos trabalhos utilizaram a implementação de um modelo vetorial denominado SMART. Segundo Ferneda (2003), o sistema SMART continua sendo uma referência no desenvolvimento de sistemas de recuperação de informação, e ainda é utilizado para pesquisas em ambiente acadêmico.

> O projeto SMART (*Sistem for the Manipulation and Retrieval of Text*) teve início em 1961 na Universidade de *Harvard* e mudou-se para a Universidade de Cornell após 1965. O sistema SMART é o resultado da vida de pesquisa de Gerard Salton e teve um papel significativo no desenvolvimento de toda a área da Recuperação de Informação. No sistema SMART cada documento é representado por um vetor numérico. O valor de cada elemento desse vetor representa a importância do respectivo termo na descrição do documento. Estes pesos podem ser atribuídos manualmente, o que necessitaria de pessoal especializado trabalhando durante certo tempo. No entanto, o sistema SMART fornece um método automático para o cálculo dos pesos não só dos vetores que representam os documentos, mas também para os vetores das expressões de busca (FERNEDA, 2003).

• Modelo probabilístico: conforme seu nome, o modelo probabilístico é baseado na teoria matemática das probabilidades.

> Esse modelo supõe que exista um conjunto ideal de documentos que atende a cada uma das possíveis buscas que podem ser feitas no sistema. A partir do primeiro conjunto de documentos resultantes de uma busca, o usuário seleciona alguns que considera relevantes para responder à sua necessidade de informação. A expressão de busca, juntamente com os documentos que foram selecionados como relevantes, é submetida novamente ao sistema de informação, procurando refinar a busca e tentando aproximar-se cada vez mais do conjunto ideal de documentos. Este processo interativo é conhecido como *Relevance Feedback* (SOUZA et al., 2006).

A principal virtude do modelo probabilístico está em reconhecer que a atribuição de relevância é uma tarefa do usuário (FERNEDA, 2003). Interessante notar que este modelo reutiliza o resultado da primeira busca para refinar ainda mais seus resultados. Ferneda (2003) relata, entretanto, que a complexidade do modelo desencoraja muitos desenvolvedores de sistema a abandonar os modelos booleano e vetorial.

• Modelo *fuzzy*: Shaw (1999) afirma que este modelo está baseado na lógica *fuzzy*, cujo objetivo é capturar e operar com a diversidade, a incerteza e as verdades parciais dos fenômenos da natureza de uma forma sistemática e rigorosa. O modelo

f*uzzy* é baseado no mundo real, no qual a imprecisão e a incerteza são intrínsecas à recuperação de informações (PERES; BOSCARIOLI, 2008).

> Nesse modelo busca-se estender o conceito da representação dos documentos por palavras-chave, assumindo que cada *query* determina um conjunto difuso e que cada documento possui um grau de pertencimento a esse conjunto, usualmente menor do que 1. O grau de pertencimento pode ser determinado pela ocorrência de palavras expressas na query, tal como no modelo booleano, mas pode também utilizar um instrumento – como um tesauro – para determinar que termos relacionados semanticamente aos termos índice também confiram algum grau de pertencimento ao conjunto difuso determinado pela *query*. (SOUZA et al., 2006)

2.1.1.2 Modelos dinâmicos

Os modelos dinâmicos, segundo Ferneda (2003), apresentam como principal característica o reconhecimento da importância do usuário na definição das representações dos documentos. Os principais modelos dinâmicos e suas características são:

• Sistemas especialistas: segundo Ferneda (2003), um sistema especialista é um sistema computacional que procura representar o conhecimento de um especialista humano em um domínio particular, de maneira a auxiliar na tomada de decisões e na resolução de problemas relacionados a esse domínio. Este tipo de sistema já contém dados associados sobre o tema que deseja recuperar. Um sistema especialista é um programa de computador associado a um "banco de memória" que contém conhecimentos sobre uma determinada especialidade (TEIXEIRA, 1998).

> A ideia subjacente à construção dos sistemas especialistas é que a inteligência não é apenas raciocínio, mas também memória. É comum considerarmos inteligente uma pessoa que possui grande quantidade de informação sobre um determinado assunto. Assim, os sistemas especialistas obedecem ao princípio de que memória é condição necessária para a inteligência. Os sistemas especialistas fazem parte de uma classe de sistemas ditos "baseados em conhecimento", desenvolvidos para servirem como consultores na tomada de decisões em áreas restritas. Estes sistemas são adequados para a solução de problemas de natureza simbólica, que envolvem incertezas resolvíveis somente com regras de "bom senso" e com raciocínio similar ao humano. Permitem representar o conhecimento heurístico na forma de regras obtidas através da experiência e intuição de especialistas de uma área específica. A construção de sistemas especialistas obedece ao princípio de que a simulação da inteligência pode ser feita a partir do desenvolvimento de ferramentas computacionais para fins específicos (FERNEDA, 2003).

• Redes neurais: o modelo de redes neurais copia o conceito de neurônios presente na Biologia. Ferneda (2003) afirma que o neurônio é uma célula formada por três seções com funções específicas e complementares: corpo, dendritos e axônio. Estas três funções são definidas também na recuperação da informação.

> De uma forma simplificada, a recuperação de informação lida com documentos, termos de indexação e buscas. Uma tarefa comum para um sistema de recuperação de informação é pesquisar documentos relevantes que satisfazem uma determinada expressão de busca através dos termos de indexação. Pode-se dizer que em um sistema de recuperação de informação de um lado estão as expressões de busca, do outro lado estão os documentos e no meio ficam os termos de indexação. Essa estrutura pode ser vista como uma rede neural de três camadas: a camada de busca seria a camada de entrada da rede neural, a camada de documentos seria a saída e a camada de termos de indexação seria uma camada central. Não existem evidências conclusivas da superioridade das redes neurais em relação aos modelos tradicionais de recuperação de informação (FERNEDA, 2003).

As redes neurais oferecem muitas características atrativas no processo de recuperação de informação, principalmente a habilidade inata de se adaptarem às modificações nas condições do "ambiente", representado pelas buscas dos usuários (DOSZKOCS; REGGIA; LIN, 1990). Através da aprendizagem, o sistema busca gradualmente adequar os pesos das conexões, a fim de melhor representar a relevância percebida através da interação do usuário (FERNEDA, 2003). É possível perceber que neste tipo de modelo o usuário e o ambiente são relevantes, já que influenciam diretamente no processo do mesmo. O sistema aprende um pouco a cada nova interação do usuário.

• Algoritmos genéticos: segundo Ferneda (2003), a aplicação dos algoritmos genéticos na recuperação de informação representa um novo modelo para todo o processo de recuperação. Este mesmo autor cita que as representações dos documentos podem ser vistas como um tipo de "código genético". Ele ainda afirma que nesse código genético um cromossomo é representado por um vetor binário onde cada elemento armazena o valor 0 ou o valor 1, correspondendo respectivamente à presença ou ausência de um determinado termo na representação do documento.

> A aplicação dos algoritmos genéticos na recuperação de informação se apresenta apenas como uma possibilidade, uma proposição para futuras implementações de sistemas com características evolutivas. Apesar da característica evolutiva representar uma forma inovadora de abordar o problema da recuperação de informação, introduz diversos questionamentos relacionados aos efeitos de sua inerente imprevisibilidade quando utilizado em situações reais (FERNEDA, 2003).

2.1.2 Organização da informação na WEB

A folksonomia e a navegação facetada são estratégias utilizadas na recuperação da informação para *web*. Estas estratégias são mais simples de serem adotadas na Internet e podem ser vistas em inúmeros sites. A folksonomia é uma técnica difundida para viabilizar o acesso e a organização da informação. Ela é um sistema de indexação livre, em que o próprio usuário indexa, por meio de *tags*, os dados de acordo com seu ponto de vista.

Mesmo que essa técnica possa dar a sensação de que as necessidades do usuário sejam contempladas, há percepções de problemas (desvantagens) nesse sistema, sendo um deles o "caos" informativo (RODRIGUES; CRIPPA et al., 2011). Isso ocorre porque os usuários podem utilizar palavras de acordo com suas próprias regras ou habilidades, trazendo informações que nada tem a ver com o que é comumente utilizado. Um exemplo, é utilizar de um termo para lembrar de uma música, mas outro usuário utilizar-se de outro termo totalmente diferente, mas que faça sentido para ele.

Mais que isso, a classificação por meio de *tags* pode ser utilizada para posteriormente oferecer conteúdo. Velsen e Melenhorst (2009) cita um exemplo, em uma comunidade de música *online*, onde um usuário identifica alguns clipes de vídeos como o BRITPOP. A partir disso, o sistema pode inferir que o usuário é interessado em BRITPOP (assumindo que os usuários não iriam olhar para estes vídeos e marcá-los se eles não gostassem deles) e, consequentemente, recomendar os 10 vídeos mais assistidos com a respectiva *tag*. (VELSEN; MELENHORST, 2009).

Outra técnica utilizada muito na *web*, nas páginas de Internet, é a navegação facetada. Koren, Zhang e Liu (2008) cita que a pesquisa facetada está se tornando um método popular para permitir que os usuários pesquisem de forma interativa e naveguem por informações complexas. Um sistema de busca facetada apresenta aos usuários opções de refinamento de busca (KOREN; ZHANG; LIU, 2008). Este tipo de pesquisa facilita a vida do usuário permitindo encontrar o objeto que procura mais intuitivamente.

Facetas podem ajudar a encontrar o que o usuário está procurando, garantindo que você sempre obterá algum resultado. Como desenvolvedor, facetas permitem que você exponha os critérios de pesquisa mais úteis para navegar pelo seu corpo de busca. Em aplicativos de

varejo *online*, a navegação facetada geralmente é criada sobre marcas, departamentos (sapatos infantis), tamanho, preço, popularidade e classificações.

2.2 Sistemas de recuperação da informação

Dentro do tema da recuperação da informação temos os sistemas de recuperação da informação. Para Souza et al. (2006 apud LANCASTER, 1968), os SRIs são a interface entre uma coleção de recursos de informação, em meio impresso ou não, e uma população de usuários; e desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle desses; e distribuição e disseminação aos usuários. A partir deste conceito, podemos concluir que os sistemas de recuperação da informação permitem a interação do trabalho manual (humano) com o trabalho automático através de softwares.

Um dos diversos diagramas que descrevem o processo de recuperação de informação em sistemas é o de Cardoso (2000) apresentado na Figura 4, que destaca o modo em que se dá a recuperação de informação em sistemas automatizados. Os documentos são indexados e o usuário especifica a consulta.

Figura 4: Exemplo de sistema de recuperação da informação
Fonte: Gey (1992)

De acordo com Ferneda (2009) os sistemas de recuperação de informação têm por função representar o conteúdo dos documentos do corpus e apresentá-los ao usuário de uma

maneira que lhe permita uma rápida seleção dos itens que satisfazem total ou parcialmente a sua necessidade de informação.

> Já havia anteriormente apontado o fato de que os SRIs não informam o usuário – no sentido de mudar seu conhecimento sobre objeto de sua questão –, mas apenas o informam sobre a possível existência de documentos atinentes à questão, além de características desses documentos; e procura, em outro trabalho, analisar os SRIs subdividindo-os em seis subsistemas: de documentos, de indexação, de vocabulário, de busca, de interface com o usuário e de *matching*. (SOUZA et al., 2006 apud LANCASTER, 1968)

Souza et al. (2006 apud CHOWDHURY, 2004) entendem que o conceito de recuperação e informações - é como consequência, o conceito de sistemas de recuperação de informações – é auto explanatório, e divide os SRIs em subsistemas de documentos, de usuários, e de busca/recuperação; detalhando cada um desses subsistemas. Os SRIs servem de ponte entre o mundo dos criadores de informações e os usuários dessas, e para isso colecionam-nas e as organizam. (SOUZA et al., 2006 apud CHOWDHURY, 2004).

Souza et al. (2006 apud SALTON; MCGILL, 1983), definem SRIs como sistemas que lidam com as tarefas de representação, armazenamento, organização e acesso aos itens de informação.

> Sistemas de recuperação de informação organizam e viabilizam o acesso aos itens de informação, desempenhando as atividades de: Representação das informações contidas nos documentos, usualmente através dos processos de indexação e descrição dos documentos; Armazenamento e gestão física e/ou lógica desses documentos e de suas representações; Recuperação das informações representadas e dos próprios documentos armazenados, de forma a satisfazer as necessidades de informação dos usuários. Para isso é necessário que haja uma interface na qual os usuários possam descrever suas necessidades e questões, e através da qual possam também examinar os documentos atinentes recuperados e/ou suas representações (SOUZA et al., 2006).

Os sistemas de recuperação de informação são muitas vezes confundidos com os sistemas de banco de dados. Ferneda (2003) explica exatamente a diferença entre os dois sistemas.

> O processo de recuperação de informação consiste em identificar, no conjunto de documentos (corpus) de um sistema, quais atendem à necessidade de informação do usuário. Já o usuário de um sistema de recuperação de informação está, portanto, interessado em recuperar "informação" sobre um determinado assunto e não em recuperar dados que satisfazem sua expressão de busca, nem tampouco documentos, embora seja nestes que a informação estará registrada. Os sistemas de banco de dados têm por objetivo a recuperação de todos os objetos ou itens que satisfazem precisamente às condições formuladas através de uma expressão de busca. Em um sistema de recuperação de informação essa precisão não é tão estrita. A principal razão

> para esta diferença está na natureza dos objetos tratados por estes dois tipos de sistema. Os sistemas de recuperação de informação lidam com objetos lingüísticos (textos) e herdam toda a problemática inerente ao tratamento da linguagem natural. Já um sistema de banco de dados organiza itens de "informação" (dados), que têm uma estrutura e uma semântica bem definidas. Os sistemas de informação podem se aproximar do padrão que caracteriza os bancos de dados na medida em que sejam submetidos a rígidos controles, tais como vocabulário controlado, listas de autoridades, etc (FERNEDA, 2003).

2.2.1 Etapas básicas do SRI

Os sistemas de recuperação de informação contemplam duas etapas básicas: a indexação e a busca. Cada uma destas etapas será analisada nesta seção. As duas devem estar bem estruturadas para que o sistema alcance resultados de sucesso.

2.2.1.1 Indexação

Ferneda (2003) ressalta que o processo de representação, busca descrever ou identificar cada documento do corpus através de seu conteúdo, na qual é geralmente realizada através do processo de indexação. O autor ainda cita que durante a indexação são extraídos conceitos do documento através da análise de seu conteúdo e está representação identifica o documento e define seus pontos de acesso para a busca e pode também ser utilizada como seu substituto.

Para este trabalho, o conceito que melhor resume indexação é o de Brown, Fugmann e Suenonius (1977): a indexação é a operação que descreve e identifica o conteúdo de um documento, através de termos. Ele afirma ainda que os conceitos dos documentos podem ser representados por termos selecionados através da linguagem natural ou por símbolos.

Segundo Vieira (1988), a indexação é uma técnica de análise de conteúdo que condensa a informação significativa de um documento, através da atribuição de termos, criando uma linguagem intermediária entre o usuário e o documento, podendo ser realizada pelo homem (indexação manual), ou por programas de computador (indexação automática).

Interessante observar o passo a passo da indexação manual, representada por Vieira (1988), e composto por três fases, sendo todo o processo intelectual.

> Durante a indexação manual os conceitos são extraídos por um processo de análise intelectual, que compreende basicamente três fases: 1. Compreensão do conteúdo do documento, através da leitura completa do texto ou do título, do resumo e de outras partes que compõem um documento. 2. Identificação dos conceitos, estabelecendo o ambiente lógico e as diferenças entre os fenômenos, os processos, as propriedades, as operações, os equipamentos, etc. 3. Seleção dos conceitos, observando alguns fatores como: exaustividade, especificidade e consistência. (VIEIRA, 1988)

Já a indexação automática é tratada por diversos estudos na literatura. Ferneda (2003) ressalta que a automação do processo de indexação só é possível através de uma simplificação na qual se considera que os assuntos de um documento podem ser derivados de sua estrutura textual através de métodos algorítmicos. Ele ainda conclui que a principal vantagem da automação está no seu baixo custo, considerando o crescente barateamento dos computadores e dos softwares.

Ferneda (2003) afirma que os métodos automáticos de indexação geralmente utilizam "filtros" para eliminar palavras de pouca significação (stop-words), além de normalizar os termos reduzindo-os a seus radicais, processo conhecido como stemming. O autor ainda fala que essa forma de indexação seleciona formas significantes (termos ou frases) dos documentos, desconsiderando os significados que os mesmos podem possuir de acordo com os contextos. Esta abordagem pode apresentar falhas na forma de indexação, apesar da sua ampla utilização.

É preciso ressaltar o conceito e o processo de análise das stop-words. É importante a definição destas palavras para o correto funcionamento do sistema de recuperação da informação.

> O processo de remoção de *stop-words* é utilizado para remover um conjunto de palavras que aparecem com muita frequência no texto. Estas palavras, chamadas de *stop-words*, geralmente são preposições, artigos, conjunções, alguns verbos, nomes, adjetivos e advérbios. Para isto, deve ser criada uma lista, denominada *Stop-List*, no idioma em que se está trabalhando, contendo estas palavras consideradas irrelevantes. Este processo faz-se necessário para retirar do texto palavras que não tem nenhuma importância, diminuindo assim o tamanho das estruturas de indexação e facilitando a mineração (BARION; LAGO, 2015).

Já para Silva e Fujita (2012), o objetivo da indexação é o de representar o conteúdo informacional do documento, tendo em vista sua recuperação, para tanto, realiza-se um exame do documento a fim de identificar conceitos pelos quais a tematicidade de um documento estará representada. O autor ainda ressalta que essa tematicidade é determinada pelo indexador através da leitura do documento, tendo em mente as necessidades informacionais da comunidade usuária do sistema de informação.

2.2.1.2 Busca

Finalizada a discussão quanto à indexação, parte-se para a etapa de busca. Ferneda (2003) ressalta que no centro do processo de recuperação de informação está a função de busca, que compara as representações dos documentos com a expressão de busca dos usuários e recupera os itens que supostamente fornecem a informação que o usuário procura. A questão da relevância também é primordial no retorno dos resultados. Ferneda (2003) cita que o fato de um termo utilizado na expressão de busca aparecer na representação de um documento não significa que o documento seja relevante para a necessidade do usuário.

A busca é a forma que o usuário possui para encontrar o que deseja naquele momento. A necessidade de informação do usuário é representada através de sua expressão de busca, que pode ser especificada em linguagem natural ou através de uma linguagem artificial, e deve resultar na recuperação de um número de documentos que possibilite a verificação de cada um deles a fim de selecionar os que são úteis (FERNEDA, 2003).

Barion e Lago (2015) citam que a recuperação da informação é feita através de uma entrada do usuário, ou seja, através de uma consulta para que os documentos relevantes sejam encontrados. Os processos de Recuperação da informação geralmente se baseiam em buscas por palavra-chave ou busca por similaridade (HAN; KAMBER; PEI, 2011).

Ferneda (2003) ressalta assertivamente que a principal dificuldade do usuário está em predizer, por meio de uma expressão de busca, as palavras ou expressões que foram usadas para representar os documentos e que satisfarão sua necessidade. O usuário poderá ter dificuldades ao utilizar a busca de um sistema, ressaltando a importância de uma boa estratégia na formulação da busca.

Ferneda (2003) também questiona a diferença entre sistemas de banco de dados e sistemas de recuperação de informação.

> Com o aumento da quantidade de documentos disponibilizados nos sistemas de informação este processo de predição, que nunca é tão preciso como nos sistemas de banco de dados, é dificultado pelo número elevado de documentos resultantes das buscas. Assim, não é suficiente predizer um ou mais termos utilizados para indexar os documentos desejados, é necessário também evitar a recuperação de documentos não relevantes, minimizando o esforço em verificar a relevância de tais documentos (FERNEDA, 2003).

Silva e Fujita (2012) concluem que a indexação deve ser feita tendo em vista o conhecimento prévio do indexador, as necessidades informacionais do usuário, a política de indexação da unidade de informação e a estrutura textual dos documentos. Ferneda (2003) também conclui que a eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que o mesmo utiliza. Um modelo, por sua vez, influencia diretamente no modo de operação do sistema.

2.2.2 Sistema de recomendação

A maior parte dos sistemas de recuperação de informação desenvolvidos hoje em dia é concebida para atender as necessidades de um usuário padrão. A recomendação adequada de um filme, por exemplo, pode fazer a diferença entre conquistar o usuário ou perdê-lo. Devido a esta necessidade de conquista, destacam-se dentro dos sistemas de recuperação, os sistemas de recomendação. Estes têm se apresentado como um fator facilitador no momento de "cativar" o usuário.

Os Sistemas de Recomendação auxiliam no aumento da capacidade e eficácia deste processo de indicação já bastante conhecido na relação social entre seres humanos (RESNICK; VARIAN, 1997). Segundo Reategui e Cazella (2005), sistemas de recomendação podem ser definidos como sistemas que procuram auxiliar indivíduos a identificarem conteúdos de interesse em um conjunto de opções que poderiam caracterizar uma sobrecarga. São sistemas que procuram facilitar a penosa atividade de busca por conteúdo interessante.

Gómez et al. (2015) citam que sistemas de recomendação são utilizados há muitos anos para diferentes razões, sendo muito utilizados nos setores de comércio eletrônico e de redes sociais.

Os sites de comércio eletrônico utilizam os Sistemas de Recomendação empregando diferentes técnicas para encontrar os produtos mais adequados para seus clientes e aumentar, deste modo, sua lucratividade. Atualmente, um grande número de websites emprega os Sistemas de Recomendação para levar aos usuários diferentes tipos de sugestões, como ofertas casadas ("clientes que compraram item X também compraram item Y"), itens de sua preferência, itens mais vendidos nas suas categorias favoritas, entre outros.

As estratégias de recomendação podem ser utilizadas de acordo com os requisitos que se deseja.

Cazella, Nunes e Reategui (2010) apresentam algumas estratégias utilizadas:

- Reputação do Produto: uma estratégia bastante utilizada em sistemas de recomendação é baseada no uso das avaliações dos usuários para estabelecer a reputação de um item, ou produto. Após conhecer determinado item, consultando-o ou adquirindo-o, o usuário tem a possibilidade de deixar uma avaliação sobre este.
- Recomendações por Associação: este tipo de recomendação é obtido através de técnicas capazes de encontrar em uma base de dados associações entre itens avaliados por usuários (comprados, lidos e outros).
- Associação por Conteúdo: também é possível fazer recomendações com base no conteúdo de determinado item, por exemplo, um autor, um compositor, um editor, entre outros. Para possibilitar este tipo de recomendação, é necessário que se encontrem associações num escopo mais restrito.

Vale ressaltar que a associação por conteúdo é a estratégia que será utilizada neste trabalho. Os arquivos de legendas dos filmes e séries serão utilizados para recomendação de outros filmes, bem como a classificação por gênero e a possibilidade de pesquisa dos arquivos indexados. Além das mais variadas estratégias utilizadas por sistemas de recomendação, várias técnicas de recomendação têm surgido visando à identificação de padrões de comportamento (consumo, pesquisa e outros) e utilização destes padrões na personalização do relacionamento com os usuários. Estas técnicas fundamentam o funcionamento dos Sistemas de Recomendação.

Cazella, Nunes e Reategui (2010) apresentam em seu trabalho as técnicas de filtragem de informação, filtragem baseada em conteúdo, a colaborativa, a híbrida e por último a filtragem baseada em outros contextos.

Segundo Cazella, Nunes e Reategui (2010 apud LOEB; TERRY, 1992) a demanda por tecnologias de filtragem de informação não é algo novo. Estes autores ressaltam que é preocupante no que se refere à quantidade de informação que estava sendo gerada pelos diversos tipos de sistema e recebidas pelos usuários, além de destacar que toda a atenção estava concentrada na geração da informação para suprir as necessidades do usuário, mas que também é importante se preocupar com o recebimento da informação, com o controle de processo, de recuperação e filtragem da informação para que esta alcançasse a pessoa que deveria utilizá-la.

A filtragem baseada em conteúdo tem sido estudada há vários anos. Segundo Cazella, Nunes e Reategui (2010 apud HERLOCKER; KONSTAN; RIEDL, 2000), por muitos

anos os cientistas têm direcionado seus esforços para aliviar o problema ocasionado com a sobrecarga de informações através de projetos que integram tecnologias que automaticamente reconhecem e categorizam as informações. Alguns softwares têm como objetivo gerar de forma automática descrições dos conteúdos dos itens e comparar estas descrições com os interesses dos usuários visando verificar se o item é ou não relevante para cada um (CAZELLA; NUNES; REATEGUI, 2010 apud BALABANOVIC; SHOHAM, 1997). Esta técnica é chamada de filtragem baseada em conteúdo segundo Cazella, Nunes e Reategui (2010 apud HERLOCKER; KONSTAN; RIEDL, 2000), por realizar uma seleção baseada na análise de conteúdo dos itens e no perfil do usuário (CAZELLA; NUNES; REATEGUI, 2010 apud ANSARI; ESSEGAIER; KOHLI, 2000).

Já a filtragem colaborativa foi desenvolvida para atender pontos que estavam em aberto na filtragem baseada em conteúdo (CAZELLA; NUNES; REATEGUI, 2010 apud HERLOCKER; KONSTAN; RIEDL, 2000). A filtragem colaborativa se diferencia da filtragem baseada em conteúdo exatamente por não exigir a compreensão ou reconhecimento do conteúdo dos itens (CAZELLA; NUNES; REATEGUI, 2010). Posteriormente, na filtragem híbrida, são combinados os pontos fortes da filtragem colaborativa e filtragem baseada em conteúdo visando criar um sistema que possa melhor atender as necessidades do usuário (CAZELLA; NUNES; REATEGUI, 2010 apud HERLOCKER; KONSTAN; RIEDL, 2000).

Por último, na filtragem baseada em outros contextos, Cazella, Nunes e Reategui (2010 apud MCDONALD, 2003) citam que a mudança mais importante a se desenvolver na nova geração de Sistemas de Recomendação é a devida complexidade na construção do modelo/perfil de usuário e, o uso apropriado desse modelo. Considerando Cazella, Nunes e Reategui (2010 apud PERUGINI; GONÇALVES; FOX, 2004) modelos/perfis de usuário propiciam indiretamente conexões entre pessoas possibilitando e direcionando a recomendações mais eficientes. Estes autores acreditam que perfis de usuário devem representar diferentes e ricos aspectos da experiência diária de um usuário, considerando a vida real como modelo.

Já Barth (2010) cita que sistemas de recomendação têm por objetivo recomendar itens (livros, músicas, artigos e fotos) que possam ser relevantes para o usuário. Ele também ressalta ainda que estes sistemas de recomendação fazem uso de uma estrutura chamada perfil de usuário. Um perfil de usuário consiste, principalmente, de conhecimento sobre as preferencias individuais que determinam o comportamento do usuário (BARTH, 2010).

Segundo Barth (2010) a maior fraqueza destes estudos é que eles assumem como premissa que todas as pessoas que estão envolvidas no processo pertencem a um conjunto homogêneo, caracterizando uma pessoa típica que poderá ser utilizada para desenvolver sistemas que podem ser usados por qualquer pessoa. Esta crítica é bastante assertiva, como por exemplo: as pessoas podem simplesmente realizar uma pesquisa para outra pessoa em seu próprio perfil, ou ainda executar alguma pesquisa esperando um resultado totalmente diferente da sua rotina (diferente do seu perfil). Sistemas alternativos que atendam a estas possibilidades devem ser desenvolvidos e disponibilizados.

> Existem algumas premissas para o desenvolvimento de sistemas de recomendação que fazem uso de um perfil de usuário: um sistema de recomendação que faz uso de um perfil de usuário não pode iniciar as suas atividades sem a criação de um perfil do usuário; é necessário representar o perfil do usuário (escolher uma técnica de representação); tais sistemas precisam de técnicas adequadas para gerar um perfil do usuário inicial, e; quando os usuários interagem com o sistema, eles fornecem informações sobre eles mesmos e sobre as suas atividades (BARTH, 2010).

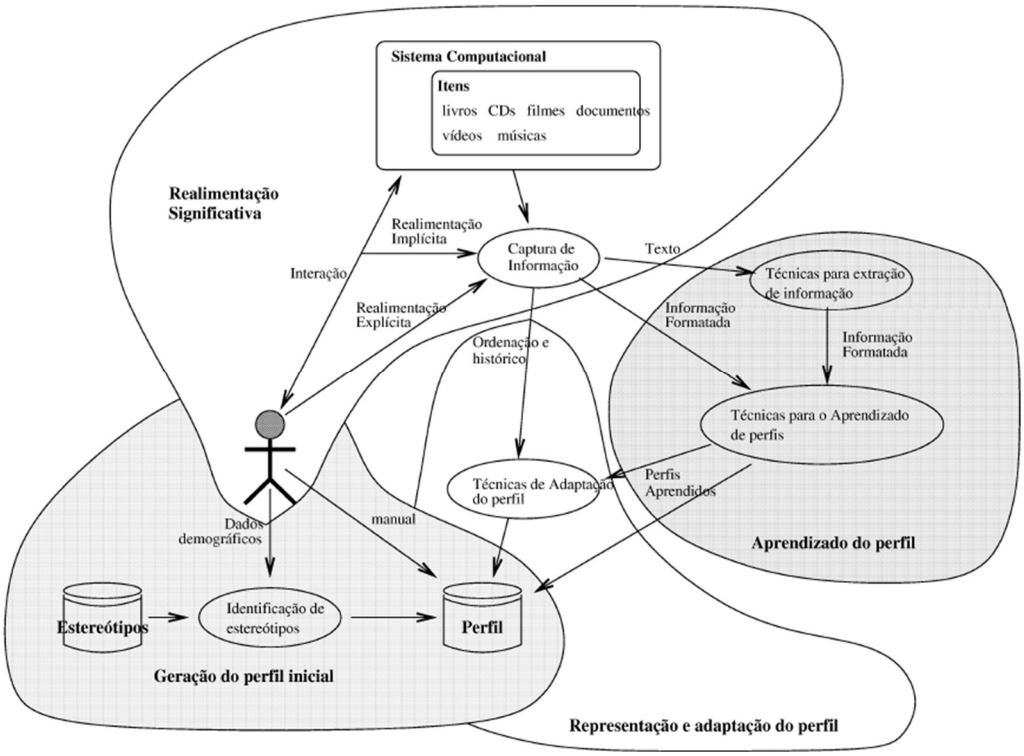


Figura 5: Criação e manutenção do perfil de usuário
Fonte: Barth (2010)

Com base nas premissas acima, é possível determinar cinco decisões de projeto (representados na figura 5) que podem ser tomadas para o desenvolvimento de módulos que permitem a criação e manutenção do perfil do usuário: a técnica de representação do perfil do

usuário; a técnica utilizada para criação do perfil do usuário inicial; a técnica de aprendizado do perfil do usuário; a fonte de realimentação relevante que representa os interesses do usuário, e; a técnica de adaptação do perfil do usuário (BARTH, 2010 apud MONTANER; LÓPEZ; ROSA, 2003). A Figura 5 mostra o relacionamento entre os módulos para que seja criado e mantido o perfil do usuário.

Por fim, Nodari (2014 apud TERVEEN et al., 1997), apresenta no quadro 1, de forma resumida, quatro formas de implementação e as caraterísticas de projeto relacionadas de sistemas de recomendação. A proposta deste trabalho mescla essas características e formas de implementação.

QUADRO 1: Sistemas de recomendação, características do projeto e formas de implementação

Características de projeto	Formas de implementação			
	Baseados em conteúdo	**Suporte a recomendações**	**Mineração de dados sociais**	**Filtragem colaborativa**
Preferências	Somente as preferências de quem busca		Coleta em sites sociais	Quem busca deve fornecer suas preferências
Papéis e comunicação	Comunicação automatizada e papéis assimétricos	O sistema apenas fornece o suporte para humanos.	Automatizado pelo sistema, potencial para comunidades. Papéis assimétricos	Automatizado pelo sistema, potencial para comunidades. papéis simétricos
Algoritmos	Aprendizagem de máquina, recuperação de informação.		Mineração de dados	Ponderação e agrupamento de preferências.
Interface de usuário	Simples	Simples	Interfaces complexas, mapas e gráficos.	Simples

Fonte: Nodari (2014 apud TERVEEN et al., 1997)

Gómez et al. (2015) citam que os recentes estudos em sistemas de recomendação preocupam mais com o alto volume de dados, ou seja, como estes bancos de dados imensos (era do big data) afetam os sistemas de recomendação, mais que isso, como eles podem ser benéficos ajudando a obter melhores resultados.

> As ferramentas e técnicas empregadas para análise automática e inteligente destes imensos repositórios são os objetos tratados pelo campo emergente da descoberta de conhecimento em bancos de dados (DCBD), da expressão em inglês *Knowledge Discovery in Databases* (KDD). Mineração de dados é a etapa em KDD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão (SILVA, 2004).

O Netflix utiliza desta enorme gama de dados também para gerar conteúdo. Hallinan e Striphas (2016) citam que House of Cards, série de maior sucesso do Netflix, foi pensada e trabalhada antes de mesmo de ir ao ar, isto é, profissionais da indústria usam algoritmos e outros recursos para orientar suas decisões sobre qual material produzir.

É necessário, entretanto, ficar atento em como empregar um investimento de centenas de milhões de reais na coleta dos dados, para no final do projeto pouca ou nenhuma informação útil ser identificada.

2.2.3 Análise de sentimentos

A análise de sentimentos é outro conceito difundido entre os sistemas de recuperação da informação. Santos et al. (2010) conceituam a análise de sentimento ou mineração de opinião como um ramo da mineração de textos preocupado em classificar textos não por tópicos, e sim pelo sentimento ou opinião contida em determinado documento.

> A análise de sentimento é uma disciplina recente que congrega pesquisas de mineração de dados, linguística computacional, recuperação de informações, inteligência artificial, entre outras. A mineração de opiniões opera sobre porções de texto de quaisquer tamanho e formato, tais como páginas *web*, *posts*, comentários, *tweets*, revisões de produto, etc. Toda opinião é composta de pelo menos dois elementos chave: um alvo e um sentimento sobre este alvo. Um alvo pode ser uma entidade, aspecto de uma entidade, ou tópico, representando um produto, pessoa, organização, marca, evento, etc. Já um sentimento representa uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo. A polaridade de um sentimento corresponde a um ponto em alguma escala que representa a avaliação positiva, neutra ou negativa do significado deste sentimento. (BECKER; TUMITAN, 2013)

Cada vez são mais comuns os estudos com análise de sentimentos. Araujo et al. (2012) apresentam uma revisão narrativa da literatura sobre técnicas computacionais utilizadas para o processo de Análise de sentimentos. Araujo et al. (2012) citam que o estudo foi realizado com artigos publicados no período de 2009 a 2011. O quadro 2 mostra os artigos encontrados e os diferentes objetivos estudados.

QUADRO 2: Artigos encontrados nas bases de dados PubMed, ISI, ACM e IEEE sobre análise de sentimento, seus objetivos e resultados

Artigo/Veículo de Publicação	Base de Dados	Objetivo	Resultados
CHEE et al., 2009. Social Visualization of Health Messages. /System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on	IEEE	Estudar temas de visualização social de sites com informações de pacientes para uma melhor exploração e uso do site.	Favoreceram o desenvolvimento de ambientes online padronizados, com o objetivo de auxiliar os usuários na busca por informações de saúde.
ZORMAN; VERLIC,2009. Explanatory approach for evaluation of machine learning-induced knowledge. / The Journal of International Medical Research	PubMed	Criar uma etapa de pré-avaliação para oferecer suporte à investigação de conhecimentos na área da saúde.	A pré-avaliação demonstrou ser útil para processar as regras extraídas, fornecendo novos padrões mais concisos e maior apoio ao especialista na avaliação dos documentos.
CHEW,EYSENBACH, 2010. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. / PLoS ONE	ISI	Analisar conteúdos de mensagens divulgadas no Twitter, durante o surto de H1N1 em 2009.	Verificou que *tweets* podem ser usados para análise de conteúdos em tempo real e fontes de opiniões e experiências. Fornecendo assim um auxilio as autoridades de saúde pública a captarem com maior facilidade às preocupações do público.
VYDISWARAN et al., 2011. Gauging the Internet Doctor: Ranking Medical Claims based on Community Knowledge. / Workshop on Data Mining for Medicine and HealthCare	ACM	Verificar a viabilidade de avaliar automaticamente a confiabilidade de uma reclamação médica baseada no conhecimento da comunidade.	Ampliou a noção de fidelidade a um site a partir de uma pontuação de confiança. Mostraram que é viável pontuar a confiabilidade de reclamações baseadas nas opiniões expressadas pelos usuários em fóruns.
YU, 2011. The emotional world of health online communities. / iConference '11	PubMed	Medir o tipo de emoção expressada por membros de comunidades online de saúde.	A análise auxiliou a compreensão das comunicações emocionais nas comunidades de saúde, definindo os participantes e comparando seus status emocionais.

Fonte: Araujo et al. (2012)

Souza (2012) cita que as soluções de análise de sentimentos foram aplicadas, na literatura, a diversos problemas: desde mineração de opiniões sobre um determinado produto em blogs e fóruns; análise automática de resenhas de filmes em sítios como IMDb; de resenhas de produtos em sítios como o Amazon; até auxílio a Sistemas de respostas a perguntas e extração de informação.

Guerra et al. (2011) citam que tradicionalmente os algoritmos de análise de sentimento (também conhecido como mineração de opinião) foram desenvolvidos para cenários estáticos e bem-controlados que têm como alvo a análise de comentários de produtos e serviços. Guerra et al. (2011) complementam que nesses cenários, as listas de palavras pré-definidas positivas e negativas (ou seja, léxicos) e tradicionais técnicas supervisionadas de aprendizado de máquina têm sido bastante bem sucedida.

Melville, Gryc e Lawrence (2009) citam que a explosão de conteúdo gerado pelo usuário na web tem levado novas oportunidades e desafios para as empresas, que estão cada vez mais preocupados com o monitoramento a discussão em torno de seus produtos. Estas informações podem gerar informações úteis sobre como melhorar produtos ou comercializá-los de forma eficaz. Melville, Gryc e Lawrence (2009) citam que um importante componente de tal análise é caracterizar o sentimento expresso em blogs sobre marcas e produtos específicos.

Melville, Gryc e Lawrence (2009) ressaltam ainda que a análise de sentimento centra-se na tarefa de automaticamente identificar se uma parte do texto expressa uma opinião negativa ou positiva sobre o assunto, já alguns estudos recentes tratam o problema como classificação de texto, utilizando uma classe de palavras.

Reis et al. (2012) confirmam que a diversidade de métodos e técnicas resulta e/ou tem impactos na diversidade de categorização de sentimentos utilizados em cada um dos métodos, nas suposições feitas e nos conjuntos de dados utilizados para validação. Reis et al. (2012) citam que grande parte destes recursos é disponibilizada apenas para a língua inglesa. Este trabalho apresenta um modelo para classificação de filmes por sentimento, utilizando termos da língua portuguesa.

Uma das aplicações da análise de sentimentos é na música digital. Os tradicionais sistemas de filtragem baseada em conteúdo e filtragem colaborativa ainda são constantemente utilizados. Na música também é possível utilizar os sistemas de recomendação. Com a análise de sentimentos o sistema pode tentar identificar o tipo de humor de uma música para recomendar ao usuário. Segundo Karmaker et al. (2015), hoje o ouvinte de música enfrenta vários obstáculos na busca de música para um contexto específico. Karmaker et al. (2015) também citam que com a extensão da biblioteca de música digital, a cada dia, está ficando mais complicado classificar ou categorizar a música acordo com as especificações emocionais do usuário.

> A necessidade de ferramentas de classificação de música está se tornando cada vez mais aparente. Ouvintes agora precisam de ferramentas para classificar a música baseada em seu humor, desfrutando da música de maneira nova e emocionante. Com o desenvolvimento da tecnologia de música digital, é essencial desenvolver um sistema de recomendação de músicas para os usuários. Já existem trabalhos realizados para a recomendação de música personalizada com base na preferência dos usuários. Estes trabalhos utilizam duas abordagens principais: filtragem baseada em conteúdo e filtragem colaborativa. Ambas as abordagens são recomendação baseadas nas preferências dos usuários observada a partir do comportamento de audição. Muitos dos sistemas de recomendação de música atual seguem a regra de marcação social, em vez de recurso extração de música. Embora estes sites sociais de marcação sejam úteis, eles não podem representa as preferencias atuais do próprio usuário (KARMAKER et al., 2015).

Mohammad e Turney (2010) citam que a criação de uma playlist de música automática classificadas por emoções poderia ser muito melhor que as padronizadas por gênero. Esse autor exemplifica que até mesmo o clima pode afetar sua emoção e seu humor, como, por exemplo: em tempo chuvoso um ouvinte pode querer ouvir uma música triste; enquanto outro ouvinte quer uma música feliz.

2.2.4 Trabalhos relacionados

Nesta seção são discutidos os trabalhos relacionados utilizados na construção desta dissertação. O primeiro trabalho relacionado é o de Barreto (2011), no qual apresenta o título: Anotação automática e recomendação personalizada de documentários brasileiros – Sistema DocUnB. O autor apresenta como problema a dificuldade de localização de documentários, resultado da crescente

produção dos mesmos, provocando a invisibilidade de grande parte da produção audiovisual brasileira, sobremaneira os filmes de cunho documental. Esta situação sugere então estudos de viabilidade do uso de mecanismos de disseminação de informações sobre este material de forma a revitalizar repositórios e coleções de vídeos, criando oportunidades para o aproveitamento contínuo de um importante acervo cultural, histórico e antropológico.

Barreto (2011) traz como proposta promover o encontro seletivo e personalizado entre espectadores e documentários, tentando alcançar este objetivo por meio de um sistema automático de recomendação de vídeos. Ele utiliza um software para retirar as falas dos personagens de cada documentário, construindo um grupo de palavras-chave que identificam o filme. Em seguida construiu também palavras-chave identificando alguns sites da Internet. Através da mescla dos resultados ele tenta construir um sistema personalizado de recomendação de documentários para o respectivo autor e usuários do site. Ele chega a concretizar um protótipo, denominado sistema DocUnB, para testar os mecanismos de recomendação.

O trabalho se assemelha com a proposta desta dissertação, entretanto o foco deste estudo é a localização de novos filmes e séries dentro de um acervo disponível. A recomendação de documentários se assemelha a recomendação de filmes aproximando os usuários do que desejam assistir. Uma diferença importante entre os trabalhos é o corpus utilizado na pesquisa: Barreto (2011) utiliza um software para retirar as falas de cada documentário. Este trabalho utiliza os arquivos de legendas que contém as falas dos personagens, ou seja, o conteúdo já está pronto, e é fiel ao que se passa nos filmes, sem erros ou equívocos, bastando apenas formatá-lo.

O resumo do estudo de Barreto (2011) é apresentado em seguida:

> Apresentação de um sistema para a Recomendação Personalizada de vídeos em Arquivos Audiovisuais, viabilizado em um modelo capaz de obter índices de conteúdo em entidades multimídia e compará-los aos conteúdos de sites da Internet, de forma a direcionar a visualização de filmes documentários. O sistema implementa agentes de *software* com capacidade de prospecção e decisão no âmbito da Internet que, a partir de interações com usuários humanos, podem construir uma identidade e agir seletivamente para a difusão de informações sobre vídeos. Se pretende facilitar a circulação de documentários, que, de outra forma, poderiam estar limitados a exibições para um público restrito e não necessariamente atento ao conteúdo específico dos filmes. Utilizando o contexto profissional acadêmico para determinar preferências quanto a materiais audiovisuais, foi realizada a recomendação baseada no conteúdo de filmes selecionados do acervo da Universidade de Brasília. O trabalho explora a difusão de filmes em interfaces interativas de rede e a personalização da recomendação em comunidades virtuais, com o objetivo de aumentar a significância dos documentos, por meio da análise de audiovisuais e de links, de forma a ampliar os pontos de acesso a documentários brasileiros (BARRETO, 2011).

O segundo trabalho relacionado é o de Yahiaoui (2003), o qual apresenta o título: Construction automatique de résumés vidéos - Proposition d'une méthode générique d'évaluation. Este estudo busca construir resumos de vídeos demonstrando a parte mais relevante para os usuários. O autor

mescla a utilização de várias técnicas, tanto textuais como visuais para construção dos resumos. Ele busca identificar alguns quadros dos vídeos, além das palavras-chave de cada vídeo, para montar o resumo geral do vídeo. O estudo inclui a indexação baseada no conteúdo textual extraído de legendas, com a seleção de frases curtas ou trechos de frases como índices, na tentativa de aumentar a precisão semântica na compreensão da temática principal do vídeo analisado.

Yahiaoui (2003) realiza experiências em cima do conteúdo do vídeo bem como das legendas dos vídeos. O trabalho se assemelha na tentativa de identificar as partes mais relevantes de cada vídeo, entretanto é diferente na forma de como gera o resultado, isto é, o autor gera um vídeo com imagens (quadros) alimentadas também com as frases mais importantes do vídeo. Já neste trabalho em questão são identificadas as palavras-chave, mas para descobrir novos títulos parecidos.

O resumo do estudo de Yahiaoui (2003) é apresentado em seguida:

> O rápido crescimento de documentos multimídia requer o desenvolvimento de várias ferramentas para manipulação. A criação de resumos de vídeo automáticas é uma ferramenta poderosa para a resumir o conteúdo geral do vídeo e apresentar apenas as partes mais relevante. Através desta tese, propomos uma nova abordagem para a construção e avaliação automática de resumos de vídeo. Esta abordagem baseia-se num princípio chamado de "princípio do reconhecimento Máxima". Este processo permite construir o melhor resumo para ajudar o usuário na tarefa de identificação. O melhor resumo é aquele que maximiza o número de acertos fornecidos pelo usuário. Esta técnica pode ser utilizada para sumarização de diferentes tipos de mídia. Primeiro apresentamos resumos de um primeiro método de construção de vídeos usando informações apenas visual, então estudamos vários outros métodos de multi-construção de vídeos. Então nós adaptamos desse princípio para a construção de sínteses baseiam-se exclusivamente em informação textual. Finalmente, propusemos um processo conjunto de otimização da informação visual e textual (YAHIAOUI, 2003).

O terceiro trabalho relacionado é o de Silva (2012), que apresenta o título: "Descoberta automática de temas utilizando legendas". Silva (2012) tem como objetivo desenvolver algoritmos capazes de descobrir automaticamente o tema de uma conversa. Além desse objetivo principal, há outras particularidades das legendas que podem ser analisadas e que diferenciam as séries de TV.

> Estes algoritmos podem ser usados por profissionais da indústria cinematográfica para pesquisar e visualizar cenas que compartilham algumas características, ou para produzir uma descrição concisa e detalhada de um filme, o que poderia ser um valioso contributo para um sistema de recomendação. Outro domínio de aplicação deste sistema é o anúncio contextual. A análise semântica das cenas fornece uma poderosa ferramenta para colocar anúncios relacionados nos documentos de vídeo (SILVA, 2012).

Silva (2012) conclui que a utilização das legendas é uma excelente contribuição para identificar temas nos vídeos, além de apresentarem informações importantes para o usuário. Isso foi demonstrado no estudo através da realização de testes, em que a partir de uma nuvem de palavras os

usuários identificavam as séries em questão. Foi possível também criar grupo de palavras correspondentes a series diferentes, mas que possuem o mesmo tema.

O trabalho de Silva (2012) assemelha na forma em como os temas tentam ser identificados, através das legendas, e na linguagem de programação utilizada, o Java, mas não utilizam o mesmo software para identificar as palavras-chave. Silva (2012) utiliza o

WordNet enquanto este trabalho utiliza o EXATOIP.

O resumo do estudo de Silva (2012) é apresentado em seguida:

> Este trabalho insere-se no *projecto* VIRUS (*Video Information Retrieval Using Subtitles*). O projeto VIRUS tem como *objectivo* o desenvolvimento de um sistema de Recuperação de Informações Vídeo que irá funcionar em bibliotecas de vídeos compostas por documentos legendados. Contrastando com projetos anteriores, limitamo-nos a processar filmes e séries de televisão para as quais as legendas estão disponíveis. Aspectos diferenciais deste projeto incluem a recuperação de informação com base na análise simultânea de três fluxos de informação: sinal de vídeo, legendas e sinal de áudio. O sistema permite visualizar vídeos de forma significativa e aceita consultas do utilizador para encontrar partes dos documentos de vídeo às quais correspondem as consultas. Os domínios de aplicação de um sistema como este são vastos. Pode ser usado por profissionais da indústria cinematográfica para aceder e visualizar cenas que partilham algumas características, ou para produzir uma descrição concisa e detalhada de um filme, o que poderia ser um valioso contributo para um sistema de recomendação. Outro domínio de aplicação deste sistema é o anúncio contextual. A análise semântica das cenas fornece uma poderosa ferramenta para colocar anúncios relacionados nos documentos de vídeo. O trabalho DESCOBERTA AUTOMÁTICA DE TEMAS UTILIZANDO LEGENDAS explora um dos fluxos de informação que se pretende abordar no *projecto* VIRUS, as legendas. O seu objetivo é desenvolver algoritmos capazes de descobrir automaticamente o tema de uma conversa e sugerir quais os temas mais relevantes. Além desse *objectivo* principal, há outras particularidades das legendas que podem ser analisadas e que diferenciam as séries de TV. Os textos usados foram legendas de séries como o 24hrs, Anatomia de *Grey*, Os Sopranos, e muitas outras. O trabalho foi desenvolvido em Java e os resultados que obtemos são apresentados na interface *web* do *MovieClouds*, o protótipo do *projecto* VIRUS. Apesar do projeto ainda não estar terminado, concluímos, através de testes com utilizadores que o processamento das legendas são uma excelente contribuição para identificar temas nos vídeos (SILVA, 2012).

Muitos trabalhos distintos sobre o tema de recuperação da informação baseado em conteúdo também foram encontrados na pesquisa realizada, relacionado com outros trabalhos de sistemas de recomendação de conteúdo. Dentre estes trabalhos podemos ressaltar a dissertação de Caritá et al. (2008) que buscou implementar e avaliar um sistema de gerenciamento de imagens médicas com suporte a recuperação baseada em conteúdo, além de busca de imagens por similaridade. Destaca-se o resultado das avaliações da recuperação por similaridade que demonstraram que o extrator escolhido possibilitou a separação das imagens por região anatômica. Este trabalho tenta identificar a similaridade entre os documentos (imagens) para poder possibilitar ao pesquisador encontrar casos parecidos para que ele avalie a melhor solução.

Outro trabalho que pode ser citado é o de Barreto (2007), trata da exposição sobre processos e métodos utilizados para a indexação e recuperação textual da informação semântica em vídeo, tendo como base a identificação e classificação do seu conteúdo visual e sonoro. A criação de ferramentas que podem permitir a pesquisa por entidades e conceitos registrados em filmes está sendo empreendida não somente por filmotecas e museus, mas também de forma intensa pelos produtores de mídia, que se preparam para oferecer conteúdo audiovisual personalizado via Internet e televisão digital. O trabalho cita que na implementação de aplicações que vão de bibliotecas digitais a sistemas de segurança, serão necessárias novas ferramentas que permitam o acesso facilitado ao conteúdo de audiovisuais.

Barreto (2007) ressalta ainda que são explorados quatro processos coordenados: extração de elementos, análise de estruturas, abstração e indexação, para a obtenção de um sistema automático de segmentação e identificação de conteúdos em qualquer tipo de vídeo. Como conclusão ele verifica que a recuperação de conteúdos em audiovisuais vem obtendo sucesso especialmente na área de reconhecimento de padrões e na identificação de imagens de cunho técnico, porém a pesquisa pela decodificação semântica de imagens, a extração automática de metadados descritivos está apenas começando, e faz parte da criação da máquina ideal, semelhante a nós mesmos.

Já Chella (2004), trata de algumas estratégias para recuperação e classificação de informações em arquivos multimídia. Entre as tecnologias estudadas o padrão MPEG-7

(*Moving Picture Expert Group*) foi adotado como base para o desenvolvimento do sistema que possibilita que arquivos de vídeo possam ser demarcados em segmentos e anotados com texto livre e texto estruturado. Os vários segmentos podem ser gravados e uma interface gráfica habilita a navegação com a visualização das anotações e do segmento do vídeo. As redes de comunicação com alta velocidade tem propiciado meios para que um grande volume de conteúdos multimídia na forma de arquivos digitais de vídeo, áudio e imagens sejam produzidos e disponibilizados na Internet. Com o aumento de arquivos disponibilizados e a facilidade de acesso o problema que se apresenta é a dificuldade de identificar e gerenciar um volume cada vez maior desse conteúdo. Neste trabalho é apresentado de forma sucinta o padrão MPEG-7, suas diversas ferramentas para a descrição de conteúdos multimídia e o desenvolvimento de um sistema para indexação e recuperação de informações de arquivos de vídeo digital.

Outro trabalho que deve ser mencionado é o de Goularte (2003), que trata do desenvolvimento de técnicas com suporte à ciência de contexto, baseadas nos padrões MPEG-4 e MPEG-7, para personalizar e adaptar conteúdo em TV Interativa. Este trabalho desenvolveu

técnicas com suporte à ciência de contexto para personalizar e adaptar conteúdo em TV interativa, permitindo que usuários possam acessar, sob demanda, programas contendo apenas assuntos de seu interesse e que o acesso possa ser realizado utilizando-se diferentes dispositivos.

Outro enfoque foi dado por Marques (2007), que analisa o uso de recomendações na recuperação de informação no domínio de perfis de usuário. É demonstrado que o uso de abordagens de recomendação é um critério essencial para a identificação de documentos relevantes em vários cenários de recuperação neste domínio. As abordagens de recomendação surgiram da necessidade de se prever os documentos mais relevantes para o usuário, problema central da recuperação de informação. As recomendações servem para auxiliar no processo de receber ou fornecer indicações, e um sistema de recomendação é um sistema de informação que auxilia o usuário a recuperar informação através da previsão de seus interesses.

Marques (2007) realiza uma pesquisa exploratória, onde foi realizado um estudo do sistema Currículo *Lattes* identificando casos onde o sistema não oferece respostas com a recuperação de informação tradicional, e que poderiam ser solucionados utilizando as abordagens de recomendação. Generalizando essas dificuldades, foi criado um modelo geral de recuperação de perfis de usuário, que pode ser aplicado na recuperação de perfis em qualquer contexto, não só na recuperação de currículos. Um sistema de recuperação de perfis de usuário em sites de relacionamento foi desenvolvido com base no modelo geral, a fim de validar esse modelo, que depois foi transposto para o Currículo *Lattes*. Com os resultados alcançados, espera-se não só contribuir com os conhecimentos da academia na arte da recuperação da informação, mas ampliar os recursos para satisfação da necessidade do usuário com o estudo de novos procedimentos, e, também, consolidar a recomendação como estratégia importante da recuperação de informação, com vistas a influenciar o desenvolvimento, sob a ótica dessa estratégia, de novas ferramentas de maior qualidade.

Já Silva, Alves e Bressan (2009), ressaltam que com o advento da TV Digital houve um crescimento do volume de programas de TV oferecidos pelas operadoras de TV, aumentando a dificuldade do usuário de selecionar conteúdo relevante. Além disso, os usuários de televisão não têm como tarefa principal a procura de informações como ocorre na Internet. Diante deste cenário, os sistemas de recomendação destacam-se como uma possível solução para este problema, contudo o contexto raramente tem sido explorado durante o processo de recomendação. Este artigo apresenta uma proposta de arquitetura sensível ao contexto para suporte a recomendação personalizada de conteúdo para TV Digital – intitulada de *PersonalTVware*.

Outro trabalho alternativo é citado por Cazella, Nunes e Reategui (2010). A maior parte das interfaces dos sistemas desenvolvidos hoje em dia é concebida para atender as necessidades de um usuário padrão. Deste modo, tais interfaces acabam negligenciando necessidades e interesses particulares de cada indivíduo. Através de métodos de personalização é possível criar uma interface diferente para cada usuário, modificando sua estrutura ou seu conteúdo de acordo com o perfil de cada um. Uma das técnicas empregadas na personalização de interfaces são os sistemas de recomendação, criados inicialmente para permitir que usuários pudessem receber conteúdo personalizado através do compartilhamento de informações.

Cazella, Nunes e Reategui (2010) citam que através do monitoramento das ações dos usuários, estes sistemas são capazes de identificar conteúdo, itens ou ações a serem recomendados de forma personalizada. A recomendação adequada de um livro, por exemplo, pode fazer a diferença entre conquistar o usuário ou perdê-lo. Devido a esta necessidade de conquista, a personalização tem se apresentado como um fator facilitador no momento de "cativar" o usuário.

Por último, podemos citar o trabalho de Corumba e Macedo (2011). Eles exaltam que participantes de listas de discussão costumam receber diariamente um grande volume de mensagens em suas caixas de correio eletrônico. Em boa parte dos casos, apenas algumas destas mensagens despertam de fato o interesse do usuário. Um exemplo deste tipo de lista é a assinatura eletrônica de sistemas de chamadas para submissão de artigos científicos a conferências e periódicos (calls-for-papers), que são de grande interesse para grupos de pesquisa, professores e estudantes que desenvolvem algum tipo de atividade científica.

Corumba e Macedo (2011) citam que a diversidade das chamadas entre linhas de pesquisa variadas dificulta o acesso às mais relevantes. O artigo descreve um serviço web que organiza de forma inteligente mensagens de call-for-papers recebidas em contas de correio eletrônico. O serviço realiza mineração do texto da mensagem e processamento KNN (k-Nearest Neighbors) para categorizar os calls-for-papers entre seis grandes áreas da computação. Experimentos utilizando uma base de testes mostraram um percentual de acerto na classificação em torno de 89%. Uma extensão desse serviço web para recomendação de calls-for-papers baseado na extração automática de informações de currículos Lattes (CNPq - Conselho Nacional de Pesquisa) de pesquisadores também é apresentada.

Já Sjöberg et al. (2014) demonstram em seu trabalho a preocupação em detectar cenas violentas em filmes ou vídeos na web. Sjöberg et al. (2014) explicam que o trabalho

nasceu de uma necessidade dos pais de verificar se aquele filme ou vídeo contêm cenas violentas para decidir se seu filho pode ou não assistí-lo.

> Além de segmentos contendo violência física, as anotações também incluem conceitos de alto nível para as modalidades de áudio e visuais dos primeiros 17 filmes de *Hollywood*. Sete conceitos visuais (“a presença de sangue”, “luta”, “presença de fogo”, “a presença de armas”, “a presença de armas brancas”,“perseguições” e “cenas sangrentas”) e três conceitos de áudio (“presença de gritos”,“tiros” e “explosões”) são fornecidos (SJÖBERG et al., 2014).

Outro modelo é apresentado por Smith, Bamman e OConnor (2013): eles tentam inferir automaticamente personagens do texto, para posteriormente identificar personagens semelhantes entre filmes diferentes. Interessante observar que uma das suas fontes de dados é a Wikipedia.

Nesta seção foram demostrandos diversos trabalhos relacionados que corroboram com a finalidade deste trabalho, demonstrando a relevância do tema em questão.

3 METODOLOGIA

A metodologia deste trabalho seguiu as diretrizes do método Design Science. Segundo, Sordi, Azevedo e Meireles (2015) a pesquisa Design Science volta-se para resolução de problemas a partir da aplicação de novos conhecimentos científicos, essencialmente pragmática. Uma das diretrizes aborda o desenvolvimento dos artefatos, que nesta pesquisa, é um modelo, permitindo soluções satisfatórias aos problemas práticos. Enquanto muitos paradigmas de pesquisa científica objetivam descobrir "o que é verdade", a ciência do Design busca identificar o "o que é eficaz" (ALAN et al., 2004).

3.1 Especificação do modelo

A especificação do modelo consiste na apresentação das ferramentas Apache Lucene e OGMA que serão utilizadas na proposta do modelo. Em seguida, aborda a especificação do modelo, que será dividido em quatro etapas: passos iniciais (requisitos); pesquisa; classificação por gênero e sentimentos; e identificação de títulos similares. O objetivo geral será delineado nos objetivos específicos e responderão à pergunta problema (figura 6).

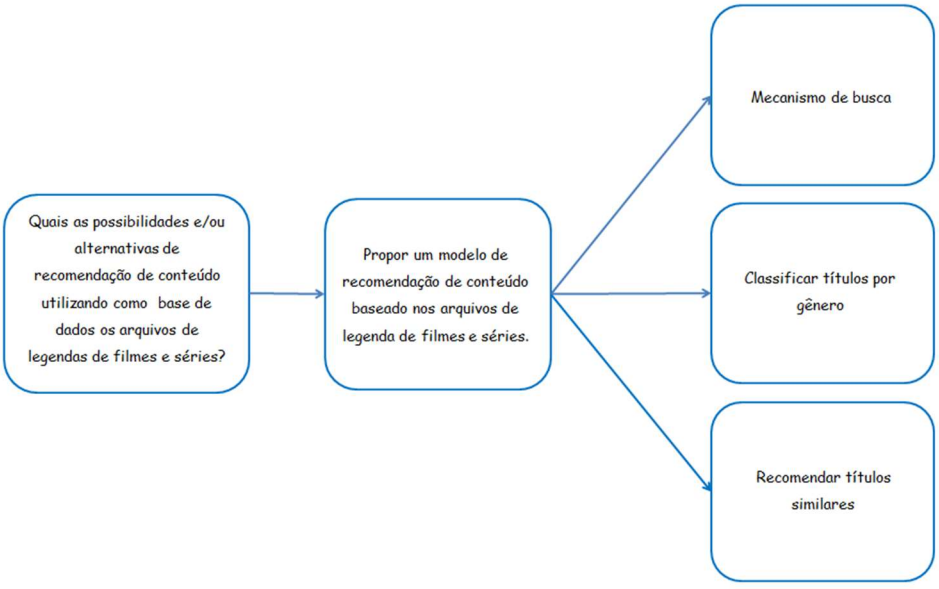


Figura 6: Organograma da pesquisa
Fonte: Elaborada pelo autor

3.1.1 Ferramentas / Softwares para construção do modelo

Antes de iniciar a descrição de cada etapa da pesquisa apresentaremos as ferramentas *Apache Lucene* e o OGMA. Os dois softwares serão utilizados na tentativa de implementação do modelo que será proposto. Esta seção visa dar transparência as ferramentas utilizadas no processo de criação do modelo, para em seguida ser construído o artefato (etapa definida pelo *Design Science*).

3.1.1.1 Apache Lucene

O Apache Lucene é uma ferramenta livre utilizada para a recuperação da informação em arquivos textuais.

> O *Lucene* é uma biblioteca de mecanismo de procura de texto altamente escalável e de software livre a partir do *Apache Software Foundation*. Você pode usar o *Lucene* em aplicativos comerciais e de *software* livre. As APIs (*Application Programming Interface*) poderosas do *Lucene* focam principalmente na indexação e na procura de texto. Elas podem ser usadas para criar recursos de procura para aplicativos, como clientes de *e-mail*, listas de correspondências, procuras da *web*, procuras de banco de dados, etc. web sites como *Wikipédia*, *TheServerSide*, *jGuru* e *LinkedIn* foram desenvolvidos com o Lucene (IBM, 2015).

Neste trabalho o Apache Lucene será utilizado para indexar os dados formatados (fases iniciais) para em seguida possibilitar a implementação da primeira etapa do modelo. A fase de análise acontece imediatamente antes de analisar a indexação e a consulta.

> Análise é a conversão dos dados de texto em uma unidade de procura fundamental, chamada de termo. Durante a análise, os dados de texto passam por várias operações: extração das palavras, remoção de palavras comuns, ignorar pontuação, redução de palavras para o formato de raiz, alteração das palavras para minúsculas, etc. A análise converte os dados de texto em *tokens* e esses *tokens* são incluídos como termos no índice do *Lucene* (IBM, 2015).

Após a indexação dos dados formatados, o Apache Lucene estará apto para atender as procuras/buscas definidas pelos usuários. O Lucene calcula a pontuação e ordena os

resultados utilizando todos os documentos que foram indexados. O site da IBM (2015) ressalta que a procura é o processo de buscar palavras no índice e de localizar os documentos que contêm essas palavras, sendo que a criação de recursos de procura usando a API de procura do Lucene é um processo direto e fácil.

Para gerar o resultado e o ranqueamento o Apache Lucene utiliza as variáveis TF (Term Frequency) e IDF (Inverse Document Frequency). Maia e Souza (2010) afirmam que a medida term frequency–inverse document frequency (TF-IDF) corresponde a uma medida estatística utilizada para avaliar o quanto uma palavra é importante para um documento em relação a uma coleção (corpus). Maia e Souza (2010) ainda ressaltam que essa importância aumenta proporcionalmente com o número de vezes em que a palavra apareça no documento e diminui de acordo com a frequência da palavra na coleção.

Maia e Souza (2013) explicam que o term frequency (TF) corresponde ao número ocorrências do termo em um documento dividido pelo número de ocorrências de todos os termos deste mesmo documento, enquanto a IDF é uma medida de grande importância para complementar a equação que avalia a importância do termo na coleção.

O Apache Lucene ainda possui as seguintes características:

> Possui algoritmos de procura poderosos, precisos e eficientes; Calcula uma pontuação para cada documento que corresponda a uma determinada consulta e retorna a maioria dos documentos relevantes classificados por essas pontuações; Suporta vários tipos de consultas poderosos, como *PhraseQuery*, *WildcardQuery*, *RangeQuery*, *FuzzyQuery*, *BooleanQuery* e outros; Suporta a análise de expressões de consulta completas digitadas pelo usuário; Permite que os usuários estendam o comportamento da procura usando classificação, filtragem e análise de expressão de consulta; Usa um mecanismo de bloqueio baseado em arquivo para impedir modificações de índices simultâneos; Permite a procura e a indexação simultaneamente; (IBM, 2015)

3.1.1.2 OGMA

O software OGMA (http://ogmaweb.com.br/ogma/) está em sua versão 1.0. O OGMA é uma ferramenta para análise de texto, cálculo da similaridade entre documentos e extração de sintagmas nominais. Sua interface é bastante simples e objetiva, conforme é demonstrado pela figura 7. Ele permite recuperar a lista ordenada dos termos mais frequentes de todos os documentos indexados, além de comparar dois documentos previamente definidos.

O principal uso do OGMA será no levantamento das palavras-chave mais relevantes de cada arquivo de legenda utilizado, excluindo as stop-words. No OGMA utilizou-se o menu “operações”, para em seguida clicar em “gerar tabelas” e por último “termos sem stop-words”.

Para construir os resultados é utilizado pelo OGMA o cálculo denominado “termo mais frequente” (TF). Conforme já foi citado, este cálculo corresponde ao número de vezes que o termo aparece no documento dividido pelo número de ocorrências de todos os termos deste mesmo documento.

Quanto maior a relevância, maior a chance de o título apresentar o que você procura. Internamente o software do modelo proposto realiza uma pesquisa em cada arquivo de legenda, do banco de dados construído, seguindo as seguintes operações:

- Quantidade de palavras do documento;
- Quantidade de vezes que a palavra-chave consta no documento;
- Divisão entre o item 1 e o item 2 para classificar por relevância (cálculo do TF);

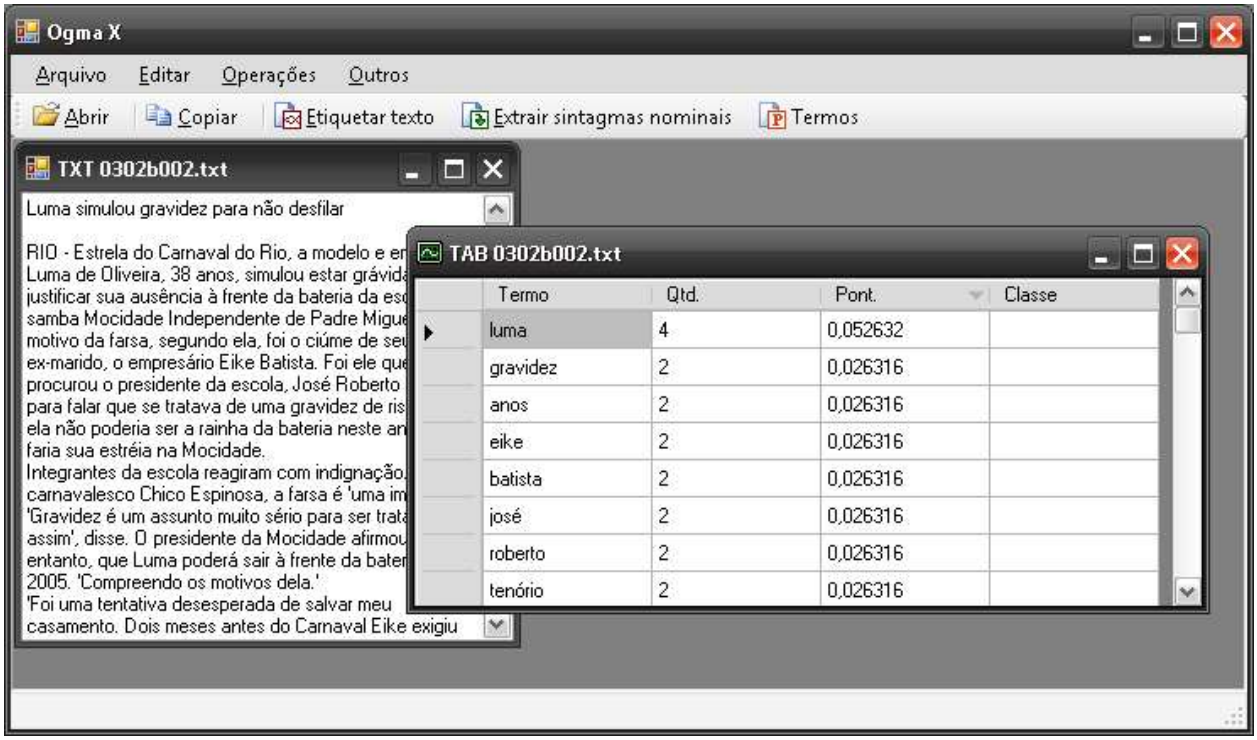


Figura 7: Interface do OGMA
Fonte: OGMA (2016)

3.2 Passos iniciais – etapa 1

Para que os objetivos específicos sejam atendidos, é necessário antes passar por alguns passos iniciais. Estes passos são comuns aos três objetivos específicos do trabalho e foram necessários para prosseguir na construção do modelo. Todos estes passos servirão como base para a construção do modelo.

Primeiramente, foi realizada a coleta dos dados (arquivos de legendas). Em seguida foi feita a formatação dos arquivos de legendas. O último passo foi realizar a indexação dos dados. A figura 8 resume estes passos com a criação do banco de dados. Cada resultado gerado em uma etapa é utilizado na próxima.

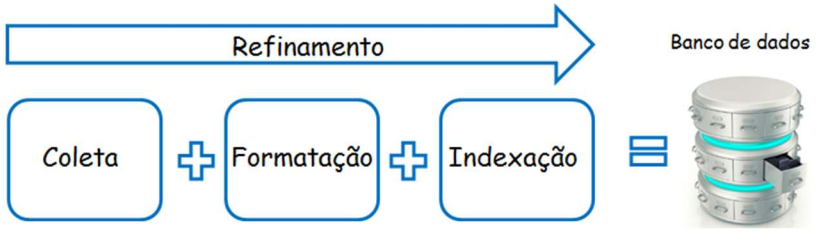


Figura 8: Requisitos do modelo - Primeira etapa
Fonte: Elaborada pelo autor

3.2.1 Coleta de dados

O primeiro passo consiste em levantar e adquirir todos os arquivos de legendas que serão utilizados. É necessário ter o maior número possível de arquivos de legendas para atender uma maior quantidade de pesquisas. No cinema, televisão e jogos eletrônicos, as legendas são o texto que acompanha uma imagem, conferindo-lhe um significado ou esclarecimento. Os arquivos de legendas mencionados são os arquivos com extensão .srt que são produzidos em diferentes línguas. Seu maior uso é na tradução de textos e diálogos de filmes, acompanhando o mesmo em sobreposição, normalmente na zona inferior da película. Não são necessários os arquivos de vídeos em questão, já que o modelo deste trabalho propõe a utilização apenas dos arquivos textos.

A seleção dos arquivos ocorreu de forma arbitrária culminando na criação de um repositório de arquivos de legendas (arquivos .srt). Para que o repositório fosse criado foi utilizado o *software opensource* Subliminal. O Subliminal é utilizado especificamente como

um sistema de busca e downloads de arquivos de legendas. O *software* foi baixado do site oficial (*http://subliminal.readthedocs.org/en/latest/*).

Foi desenvolvido um *script*, representado na figura 9, denominado *baixarlegenda.sh* para utilizar o *software* Subliminal automaticamente. Para utilizar este script foram criados dois outros arquivos texto para facilitar a inclusão de novos títulos: o primeiro com o título de todas as séries e o segundo com o título de todos os filmes que foram selecionados para *download*. O arquivo de filmes contém mais de três mil títulos e o de séries contém mais de 300 produções. Desta forma, o script executa e consulta cada linha dos dois arquivos textos criados anteriormente para pesquisar na biblioteca do Subliminal, realizando o *download* de cada arquivo de legenda. A figura 10 resume todo o processo de coleta de dados. Não serão todos os títulos que terão legendas disponíveis. Vale ressaltar que o *script baixarlegenda.sh* deve ser alterado de acordo com o arquivo texto que será utilizado.

Para atualizar a base de dados de legendas, é necessário apenas atualizar os arquivos textos, podendo ser configurado uma execução diária do *script baixarlegenda.sh*. Desta forma a base será atualizada sempre com as melhores legendas. Inicialmente foi realizado o *download* de 700 legendas para utilização neste projeto, mas o número deverá aumentar à medida que novos filmes e séries forem lançados.

```
#!/bin/bash
file=/home/armstrong/filmes.txt
while IFS= read -r line
    do
        # echo line is stored in $line
        subliminal -l pt-br -d /home/armstrong/legendas_oficiais -- "$line"
    done < "$file"
```

Figura 9: Script baixarlegendas.sh
Fonte: Elaborada pelo autor

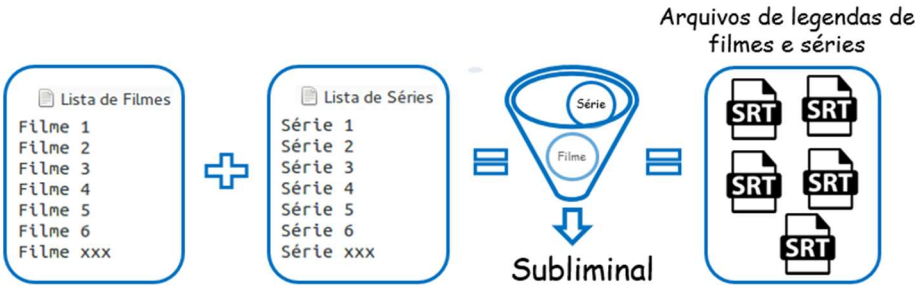


Figura 10: Etapa da coleta de dados
Fonte: Elaborada pelo autor

3.2.2 Formatação de dados

Com o repositório de legendas pronto e definido, identificou-se um problema. Os arquivos de legendas continham o tempo em que cada frase deveria aparecer. Estes números e outros caracteres especiais não poderiam ser considerados na indexação para não atrapalhar na pesquisa por conteúdo. Além disso, era necessário formatar estes dados para não atrapalhar também a construção dos termos mais relevantes, que sem a formatação trazia como resultado inúmeros números nas primeiras posições. Foi necessário preparar todos os arquivos retirando e refinando os dados que serão indexados. Finalmente, cada arquivo, deverá conter estritamente as falas dos personagens. Nenhuma outra informação, símbolo, caracteres ou imagens podem ser utilizados. Com a formatação dos dados a busca realizada pelo *Apache Lucene* será mais precisa e o resultado mais objetivo.

Para a formatação dos arquivos utilizamos o *software Notepadd++* (versão 6.8.1) que pode ser baixado em seu site oficial (*https://notepad-plus-plus.org/*). *Notepad++* é um *software* livre e editor de código-fonte, substituto do bloco de notas que suporta várias línguas. Com os arquivos de legendas todos baixados e alocados em um único diretório do sistema operacional, é aberto o *Notepadd++* para clicar em “localizar arquivos” dentro do menu “localizar”. No campo “localizar” é utilizado a expressão regular 00.+ para selecionar, na opção modo de busca, “expressão regular”. Por fim, é selecionada a pasta onde está localizado o repositório de legendas para clicar em “substituir nos arquivos”, eliminando desta forma todas as frases que começam com a informação do momento em que a legenda deve aparecer. A figura 11 apresenta a interface com a configuração efetuada. Como exemplo, pode-se perceber que as linhas 66, 71, 76, 80, 84, 88 e 92, da figura 15 foram eliminadas com este processo.

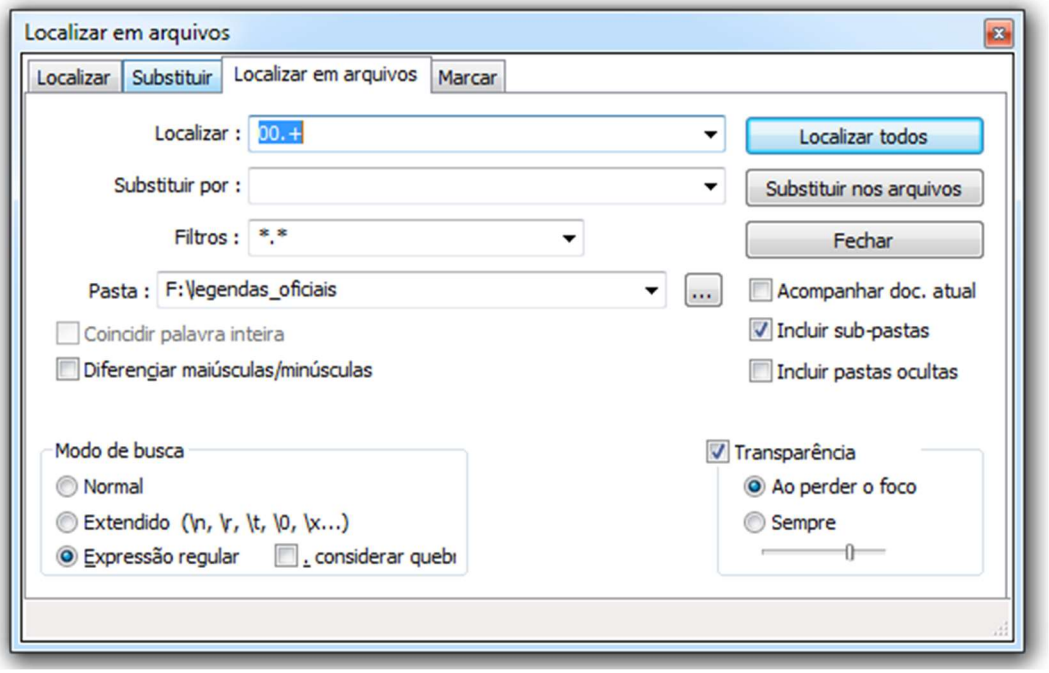


Figura 11: Utilizando a expressão regular 00.+ no notepad++
Fonte: Elaborada pelo autor

Em seguida todos os números que identificavam cada frase da legenda também foram eliminados. Para que este processo fosse realizado foram utilizados os passos anteriores dentro do Notepad++ trocando apenas o conteúdo do campo “expressão regular” para [0-9].+ eliminando as frases com números de 1 a 10. A figura 12 apresenta a interface com a configuração realizada. Como exemplo, pode-se perceber que as linhas 65, 70, 75, 79, 83, 87 e 91, da figura 13 foram eliminadas com este processo.

Estes dois processos foram realizados para todos os arquivos de legendas. Foi verificado que os acentos são ignorados na etapa de indexação realizada pelo *Apache Lucene*, ou seja, não será preciso eliminar aqui acentos e pontuações.

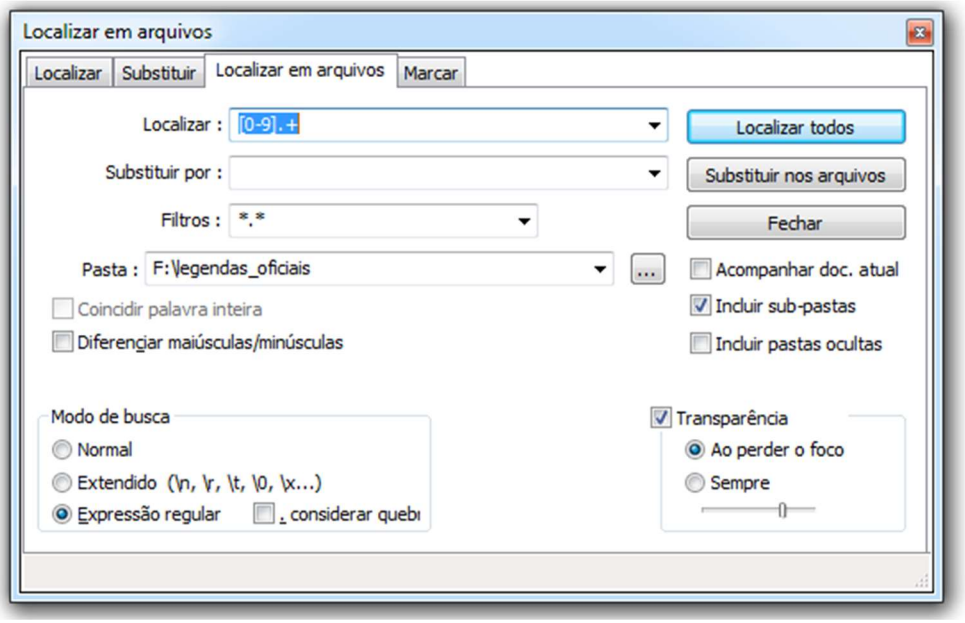


Figura 12: Utilizando a expressão regular [0-9] no Notepad++
Fonte: Elaborada pelo autor

```
14
00:06:47,650 --> 00:06:50,510
PRISÃO FEDERAL DOS EUA
- Ei, tire-me daqui!

15
00:06:50,960 --> 00:06:55,201
Hathaway, revista.
Mãos atrás das costas.

16
00:07:13,420 --> 00:07:14,850
Abram a cela 16!

17
00:07:15,750 --> 00:07:17,060
Vamos, vamos!

18
00:07:22,440 --> 00:07:24,120
Revistem tudo.

19
00:07:30,322 --> 00:07:31,955
Larguem isso!

20
00:07:40,350 --> 00:07:42,357
Um, dois, três, levanta!
```

```


PRISÃO FEDERAL DOS EUA
- Ei, tire-me daqui!


Hathaway, revista.
Mãos atrás das costas.


Abram a cela !


Vamos, vamos!


Revistem tudo.


Larguem isso!


Um, dois, três, levanta!
```

Figura 13: Comparação entre o arquivo formatado e não formatado
Fonte: Elaborada pelo autor

3.2.3 Indexação dos dados

O último passo do processo para possibilitar a construção da proposta do modelo de recuperação da informação em arquivos de legendas é a indexação. A indexação é necessária para permitir a rápida recuperação dos dados indexados, além de possibilitar a análise dos termos que mais representam cada filme ou série (item da próxima etapa). A indexação será realizada através do *Apache Lucene*. O *Lucene* é uma biblioteca de mecanismo de procura de texto altamente escalável e de *software* livre a partir do *Apache Software Foundation*. A biblioteca pode ser baixada no site oficial (*https://lucene.apache.org/*). Para este trabalho foi utilizado a versão 5.2.1 em conjunto com o *Eclipse*. O *Eclipse* é uma IDE (*Integrated Development Environment*) para desenvolvimento *Java*, porém suporta várias outras linguagens. Ele segue o modelo *open source* de desenvolvimento de *software*. A versão utilizada do *Eclipse* é a *Mars Release* (4.5.0).

O *Lucene* é utilizado para recuperar informações em arquivos. Esta funcionalidade se dá através de um motor de pesquisa, que permite a indexação de textos com alta performance. A indexação passa por um processo de análise do documento e, automaticamente, o converte para um texto simples. A extração do texto é feita a partir do objeto *Analyser*, que contém as regras para a extração do conteúdo. No entanto, é preciso saber que existem diversas implementações da classe *Analyser* que realizam essa mesma função. Neste trabalho foi utilizado a classe *BrazilianAnalyser* que contém as *stop-words* da língua portuguesa.

Na indexação automática é muito importante definir antes as *stop-words*. As *stop-words* não devem fazer parte do processo de busca, consequentemente da indexação. As *stop-words* são palavras que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido em uma busca realizada em uma *search engine*. Exemplos: as, e, os, de, para, com, sem, foi. *Stop-words* são palavras consideradas irrelevantes para a construção do índice, por isso, nada impede que você crie o seu *Analyser* com as suas próprias *stop-words*. Veja algumas palavras exemplos: “ambas”, “ambos”, “ano”, “anos”, “antes”, “ao”, “aonde”, “aos”, “apenas”, “apos” etc. Além das *stop-words* padrão, propostas pela classe *BrazilianAnalyser*, foi necessário acrescentar manualmente algumas palavras na lista de *stop-words* para ir de encontro ao objetivo proposto no trabalho. A lista de todos os termos que foram considerados como *stop-words* é apresentado no apêndice A.

Com os arquivos definidos e formatados, foi utilizado o *Apache Lucene* para indexar cada um dos arquivos. Com a indexação terminada a base de dados estará pronta para receber todas as consultas e análises necessárias, possibilitando a especificação do modelo. Os passos iniciais foram realizados para servir como base para todo o restante do trabalho.

3.3 Pesquisa por palavras-chave – etapa 2

Após o desenvolvimento dos requisitos, iniciou-se a segunda etapa do modelo. Para atender o primeiro objetivo específico foi implementado uma classe Java no eclipse para utilizar o Apache Lucene. Uma classe Java define o estado e comportamento de um objeto geralmente implementando métodos e atributos. Esta classe (apêndice B) consistiu na especificação da uma interface de pesquisa de filmes e séries já indexados na etapa anterior. A interface será composta por apenas um campo texto de um formulário na qual o usuário poderá utilizá-lo para pesquisa pelos termos que desejar. Esta interface é representada na figura 14.

O procedimento de pesquisa por palavra-chave é resumido pela figura 15. O usuário executará a pesquisa de um termo, onde a classe Java pesquisará em cada arquivo indexado no banco de dados o termo digitado. O resultado será uma lista dos documentos mais relevantes, ou seja, apresentará um ranking dos documentos que melhor representam aquele termo. A busca e o ranqueamento são realizados pelo do Apache Lucene.

Para construir o ranking é utilizado o cálculo denominado índice por peso (TF-IDF). Este cálculo beneficia termos que ocorrem bastante no documento e em poucos documentos, ou seja, quanto maior a relevância da palavra utilizada na pesquisa, maior a chance do título apresentar o que você procura.


```
13:55:29,293    INFO Buscador:45 - Total de documentos encontrados:3
13:55:29,295    INFO Buscador:46 - Tempo total para busca: 9ms
13:55:29,302    INFO Buscador:50 - Caminho:F:\legendas_oficiais\Star Trek.pt.srt
13:55:29,302    INFO Buscador:51 - Ultima modificacao:20150814
13:55:29,303    INFO Buscador:53 - Score:0.62536585
13:55:29,303    INFO Buscador:54 - --------
13:55:29,305    INFO Buscador:50 - Caminho:F:\legendas_oficiais\Enterprise.pt.srt
13:55:29,305    INFO Buscador:51 - Ultima modificacao:20150814
13:55:29,305    INFO Buscador:53 - Score:0.12264432
13:55:29,305    INFO Buscador:54 - --------
13:55:29,311    INFO Buscador:50 - Caminho:F:\legendas_oficiais\Transformers_ O Lado Oculto da Lua.pt.srt
13:55:29,311    INFO Buscador:51 - Ultima modificacao:20150814
13:55:29,311    INFO Buscador:53 - Score:0.08585103
13:55:29,311    INFO Buscador:54 - --------
```

Figura 14: Busca pelo termo enterprise utilizando o eclipse
Fonte: Elaborada pelo autor

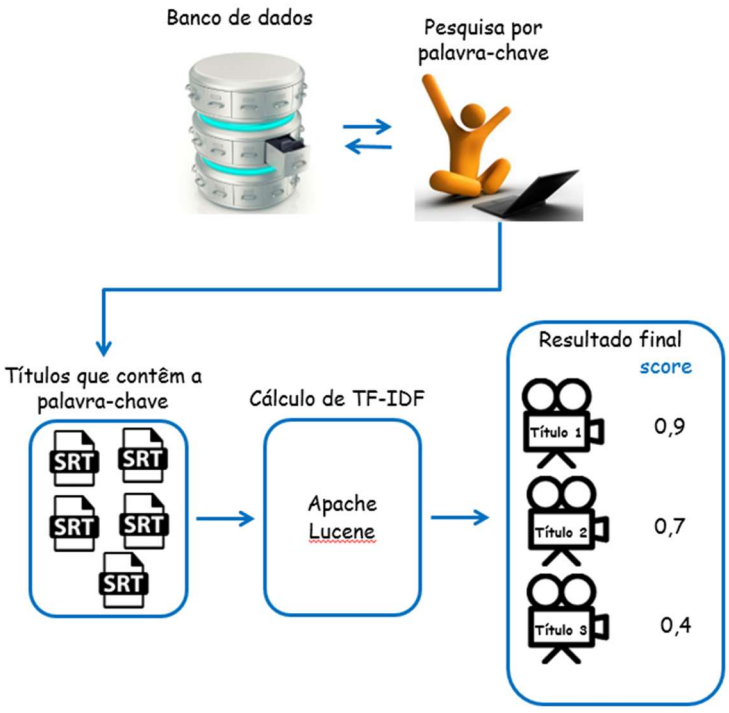


Figura 15: Proposta de modelo - Segunda etapa
Fonte: Elaborada pelo autor

3.4 Classificação por gênero – etapa 3

A terceira etapa do modelo será a especificação de uma técnica de classificação por gênero dos títulos (arquivos de legendas). A classificação por gênero ajuda o usuário a identificar de forma primária o conteúdo do filme. Pode-se considerar que é o primeiro filtro utilizado pelo usuário para que depois ele possa prosseguir e buscar mais detalhes do filme ou da série em questão. A classificação também é importante para que empresas de conteúdo tenham um modo alternativo de categorizar seus títulos, não ficando limitado ao que vêm pré-definido da direção do filme.

O passo a passo desta etapa do modelo é definido pelos itens abaixo:

- Escolher quais gêneros categorizar;
- Construção da "tabela base" por gênero;
- Utilização do OGMA para identificar as dez palavras-chave do título selecionado;
- Comparação das palavras-chave com a tabela pré-definida;
- Definição do gênero.

A construção da "tabela base" (apêndice C), composta por termos que identificarão cada gênero, é o primeiro passo desta etapa, sendo também uma das grandes dificuldades do trabalho. Esta "tabela base" servirá como referência durante todo o processo. Uma das dificuldades foi em relação às palavras que podem servir para mais de um gênero. A "tabela base" foi construída da seguinte forma:

1. Primeiro, foi necessário definir os gêneros que a "tabela base" atenderia. Os gêneros definidos, arbitrariamente, foram: Terror, Ação, Comédia, Romance, Ficção, Drama e Guerra.
2. Em seguida começou-se a identificação das palavras-chave. A principal técnica utilizada foi a seleção, por gênero, dos vinte títulos mais votados pelos usuários no site IMDb. Através destes títulos foi feita uma análise de suas respectivas palavras-chave disponibilizadas pelo site. As palavras-chave mais utilizadas foram escolhidas para fazerem parte da "tabela base". Desta forma, o primeiro esboço da tabela foi criado;
3. Ainda assim, a "tabela base" continuou sendo aprimorada a medida que mais práticas foram realizadas na construção do artefato. A cada título escolhido e suas palavras-chave identificadas, através do OGMA, mais termos relevantes eram selecionados para compor a "tabela base"; A "tabela base" foi implementada para suportar vários tipos de gêneros que poderão ser também

utilizadas em trabalhos futuros. A “tabela base” pode, por exemplo, ser utilizada para classificar músicas, por gênero, identificando o tema da mesma para recomendar aos usuários.

Gómez et al. (2015) ressaltam em seu trabalho que a escolha das palavras-chave é muito sensível nos sistemas de recomendação, e que foram necessários ajustes manuais em alguns casos durante seus testes.

O processo de classificação por gênero é resumido pela figura 16. Primeiro o usuário escolhe um título que deseje categorizar. Após a escolha do título, deve-se utilizar o OGMA para a identificação das dez palavras-chave. Dentro do *software* OGMA é selecionado o arquivo de legenda do título escolhido. Em seguida, é utilizado a opção extrair termos sem *stop-words*. O resultado é o ranqueamento de todas as palavras que compõem o arquivo de legenda. As dez primeiras palavras-chave é que serão utilizadas. Na próxima etapa é feita a relação das palavras-chave do título em questão com os termos da “tabela base” (apêndice C) criada para a classificação. O número de palavras-chave que mais estiverem relacionadas com uma única categoria classificará o título. A figura 19 apresenta como exemplo, as palavras-chave “vivos” e “medo”, logo a classificação de gênero para o respectivo filme será Terror.

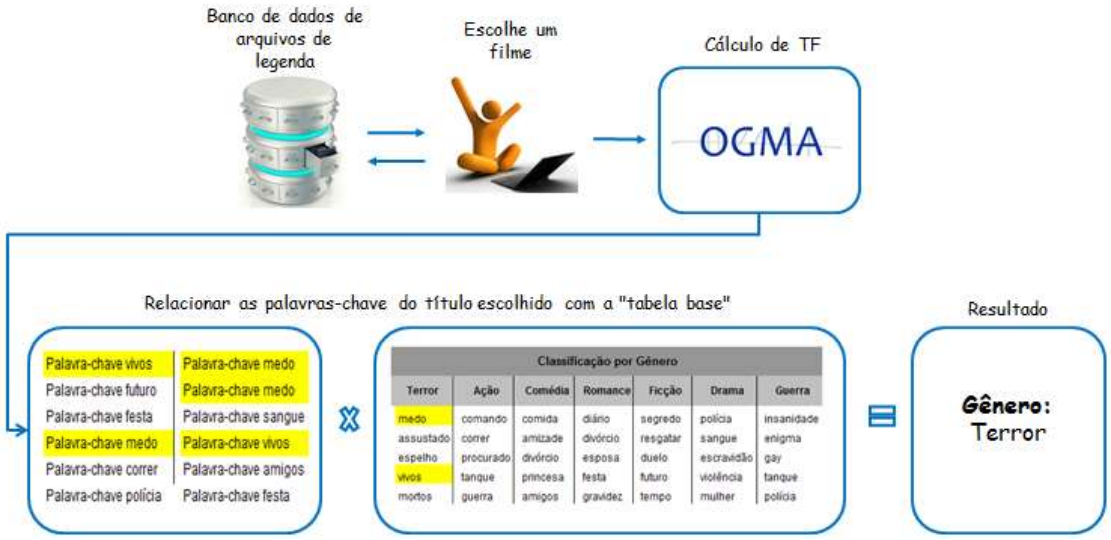


Figura 16: Proposta de modelo - Terceira etapa
Fonte: Elaborada pelo autor

3.4.1 Análise de sentimentos

Uma adaptação proposta nesta etapa do modelo é a análise de sentimentos. Cada filme pode trazer um sentimento atrelado ao mesmo. Segundo Liu (2010), análise de sentimento é o estudo de opiniões, sentimento e emoções expressas em textos. Ferreira (2010) revela que existem muitas tarefas relacionadas nessa área, como a extração de elementos do texto relacionados à opinião, a classificação da opinião quanto ao seu caráter (positivo, negativo ou neutro), comparação de sentenças quanto a suas opiniões, entre outros.

Para realizar a análise de sentimentos, os mesmos serão categorizados por palavras-chave, dando origem a "tabela base" (apêndice D) de sentimentos. Os sentimentos foram escolhidos baseado no trabalho de Karmaker et al. (2015), que citam os principais sentimentos do ser humano. São eles:

- triste
- feliz
- irritado
- tenso
- angustiado
- entediado
- cansado
- sonolento
- sereno
- satisfeito
- encantado
- animado

Depois da definição de quais sentimentos serão utilizados, começou o processo de identificação de cada termo para cada sentimento. A montagem da "tabela base" de sentimentos (apêndice D) foi baseada em dois trabalhos:

1. Tese de Osiek (2014), na qual ele propõe um modelo linguístico emocional apresentando as palavras-chave para cada sentimento;
2. Artigo de Mohammad e Turney (2010), que apresenta um método léxico, denominado

NRC *Emoticon Lexicon*, que classifica textos em 8 categorias afetivas. No site (*http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm*) é possível fazer o

download de uma planilha na qual temos as palavras definidas para cada sentimento que ele propõe;

Com a mescla destes dois trabalhos, foi possível concluir a montagem dos termos da “tabela base” de sentimentos a partir da seleção das palavras-chave para cada sentimento. A “tabela base” de sentimentos também pode ser utilizada para outros fins em trabalhos futuros. Um exemplo, é a construção de um sistema que permitirá a identificação do humor do usuário pelas redes sociais, para em seguida, possibilitar a recomendação de filmes de acordo com o respectivo humor do usuário naquele momento.

A figura 17 resume a adaptação do modelo para a análise de sentimentos. Com a “tabela base” pronta, as palavras-chave de cada sentimento serão utilizadas como parâmetro para pesquisar, no banco de dados, utilizando para isso a primeira etapa do modelo (pesquisa via *Lucene*). O resultado de cada pesquisa por palavra-chave retornará uma serie de títulos (considerados apenas os cinco primeiros por palavra-chave), por sentimento, nos quais os que aparecerem com mais frequência e com a maior soma de sua pontuação final, será o que melhor representará cada sentimento. Basicamente, será realizada a tentativa de captar qual filme ou série transmite melhor um sentimento específico.

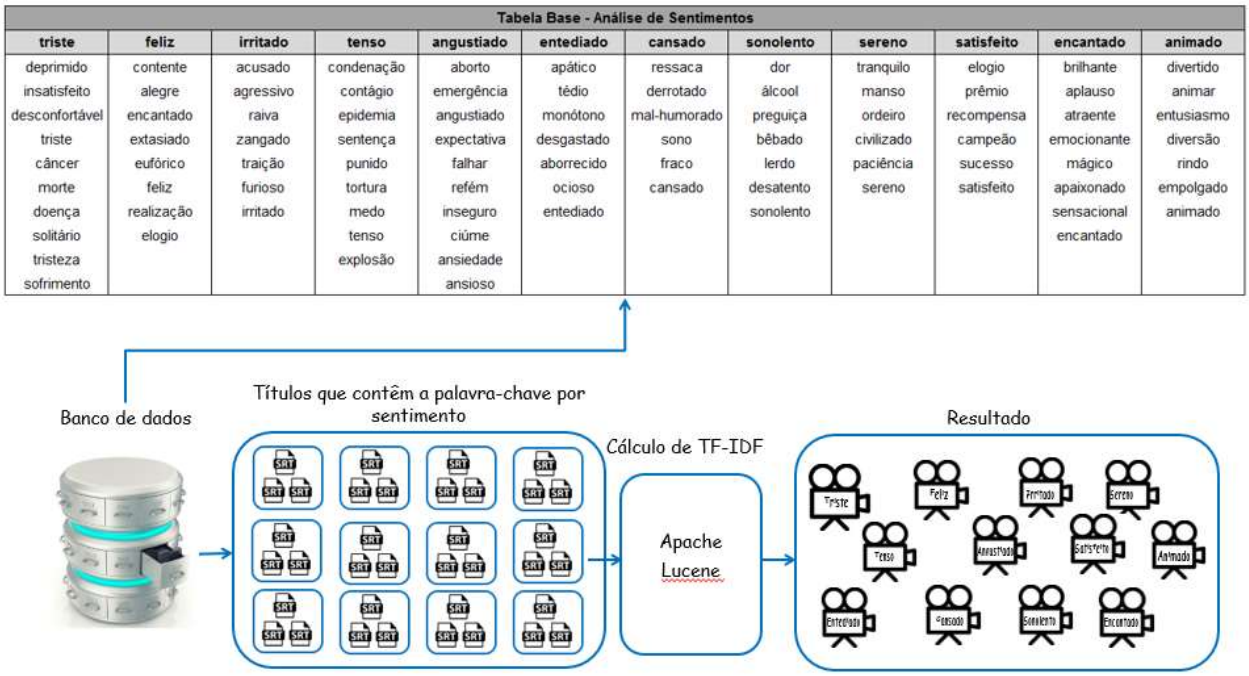


Figura 17: Proposta de adaptação do modelo - análise de sentimentos
Fonte: Elaborada pelo autor

3.5 Identificação dos títulos similares – etapa 4

A última etapa do modelo propõe a recomendação de títulos similares, a partir de um único título pré-definido. Este modelo vai propor basicamente encontrar filmes semelhantes.

A recomendação de filmes é uma abordagem muito peculiar. Um título é composto por variados temas, cenas, atores, músicas, entre outros que constroem toda a história. Os filmes podem ser completamente diferentes que mesmo assim uma única pessoa poderá gostar dos dois filmes, sendo esta a maior dificuldade e o motivo para não termos um *software* com tamanha precisão.

O passo a passo para especificação desta etapa do modelo é relacionado abaixo:

- Utilização do OGMA para identificar as cinco palavras-chave do título selecionado;
- Estas cinco palavras-chave serão utilizadas como parâmetro de pesquisa para identificar os 5 primeiros títulos do ranqueamento (via *Apache Lucene*) para cada uma;
- Os títulos recomendados serão baseados nos que aparecem com mais frequência, respeitando o título que apresentar a maior soma da pontuação final;

A figura 18 resume a etapa de identificação de títulos similares. Primeiro, são identificadas as cinco primeiras palavras-chave do título previamente escolhido pelo usuário. Para isso, será utilizado o software OGMA. Dentro do software OGMA, é selecionado o arquivo de legenda que se deseja utilizar. Em seguida, é utilizado a opção extrair termos sem stop-words. O resultado é o ranqueamento de todas as palavras que compõem o arquivo, sendo que apenas as cinco primeiras serão utilizadas.

Em seguida, cada uma das cinco palavras-chave, será utilizada como parâmetro para a pesquisa através da primeira etapa do modelo (pesquisa via Lucene). O resultado desta pesquisa levará em conta apenas os cinco primeiros títulos identificados para cada palavra-chave. O título que apresentar a maior soma da pontuação final (score) é que indicará o título semelhante.

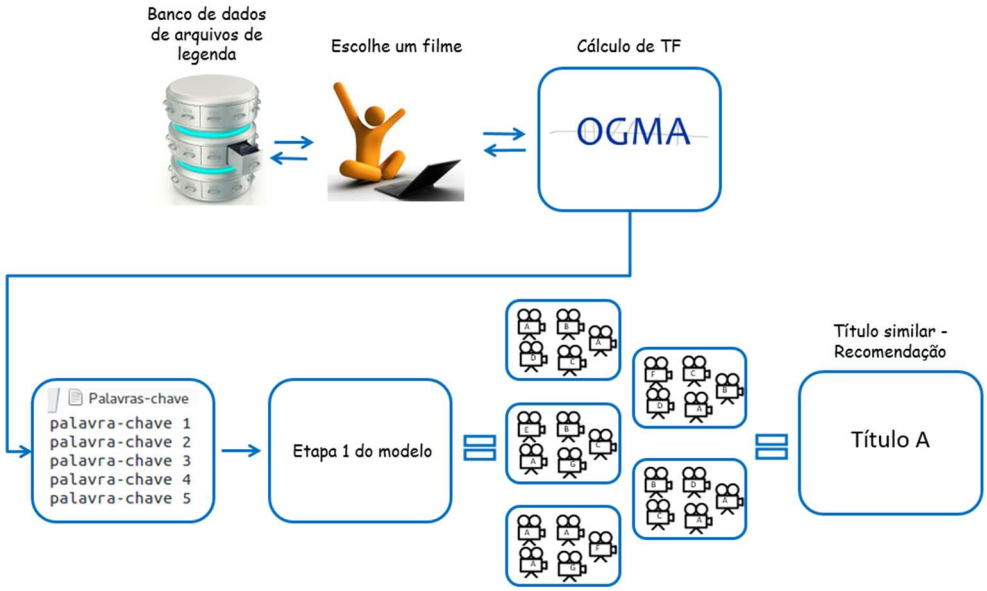


Figura 18: Proposta de modelo - Quarta etapa
Fonte: Elaborada pelo autor

3.6 Resumo do modelo

Todo o procedimento para idealização do modelo é explicado e resumido na figura 19. Após a apresentação das ferramentas utilizadas (*Apache Lucene* e OGMA), inicia-se a construção do modelo com os passos iniciais (requisitos), composto pela coleta de dados, a formatação e por último a indexação dos dados. Estes passos inicias deram origem ao banco de dados. O banco de dados é composto pelos arquivos de legendas prontos para serem utilizados por qualquer pesquisa realizada pelo usuário. O modelo foi dividido em quatro etapas para melhor visualização e entendimento.

Ainda na figura 19, é demonstrada a segunda etapa do modelo, denominada "Pesquisa". É proposto ao usuário uma abordagem mais livre, podendo pesquisar o termo que desejar, trazendo como resultado, todos os títulos que aparecem o termo pesquisado, ordenados por relevância através do *Apache Lucene*.

Em seguida, continuando a análise da figura 19, é demonstrada a terceira etapa do modelo, denominada Classificação. É uma forma inicial de se recomendar conteúdo ao usuário, direcionando o mesmo, para que em seguida ele comece uma busca mais detalhada do que queira assistir. A proposta de classificação por gênero se dá através da "tabela base" criada a partir das palavras-chave do site "IMDb". O software OGMA identificará as palavras-chave do

título escolhido para em seguida comparar com a “tabela base”. Desta forma, o gênero será definido de acordo com o maior número de termos correspondentes.

Dentro da etapa de “Classificação”, também foi realizada uma adaptação ao modelo para contemplar a análise de sentimentos. A análise de sentimentos consiste em mais uma forma inovadora de se extrair conteúdo relevante dos arquivos de legendas. A “tabela base” de sentimentos foi utilizada como parâmetro de pesquisa no banco de dados, obtendo como resultado os títulos para cada sentimento. O de maior pontuação final, dada pelo *Apache Lucene*, será o filme escolhido para representar aquele sentimento.

Por último, ainda na figura 19, é apresentada a terceira etapa do modelo, denominada “Similaridade”. Nesta etapa, é proposta a recomendação de títulos similares. Após a escolha de um título, utiliza-se o OGMA para identificar suas respectivas palavras-chave. Em seguida, é utilizada a primeira etapa modelo para pesquisa, utilizando como parâmetro as palavras-chave identificadas. O título que apresentar a maior soma da pontuação final é que será indicado como filme semelhante.

A proposta deste trabalho foi apresentar como resultado um modelo de busca e recomendação de conteúdo. Este modelo pode ser utilizado separadamente ou em conjunto, dando liberdade ao usuário. Possíveis modificações também seriam facilitadas para futuras melhorias em trabalhos futuros.

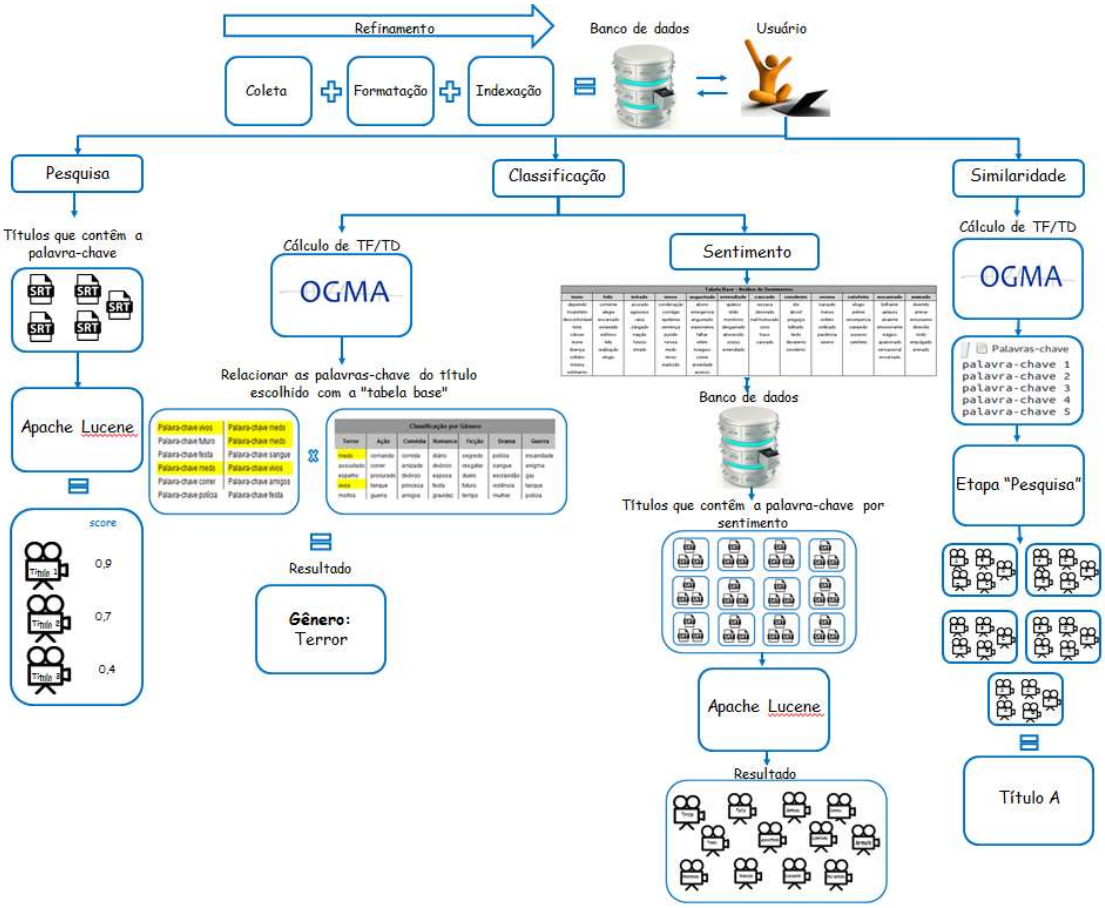


Figura 19: Overview do modelo
Fonte: Elaborada pelo autor

4 ARTEFATO: IMPLEMENTAÇÃO E TESTE DO MODELO

O desenvolvimento de artefatos, segundo os princípios da abordagem *Design Science*, é um dos meios que a academia contemporânea utiliza para responder às críticas recorrentes quanto à qualidade da produção científica: muito fragmentada, consequentemente difícil de ser aplicada a problemas concretos da sociedade, tornando-a pouco relevante (AKEN; ROMME, 2009).

O desenvolvimento do artefato será dividido em cinco seções demonstrando a utilização prática do modelo. Para testar o modelo proposto serão demonstrados alguns exemplos práticos para cada etapa do modelo. Serão demonstradas também algumas análises comparativas com o site IMDb, já apresentado neste trabalho e que hoje é referência mundial na web para pesquisas de filmes e séries. Todos as validações aqui apresentadas foram idealizadas e executadas pelo próprio autor deste trabalho.

Este capítulo inicia-se com a seção de benefícios propostos para reforçar a importância deste trabalho. Em seguida, para testar a segunda etapa do modelo, foram escolhidos no site do IMDb três títulos arbitrariamente e identificadas três palavras-chave para cada título. Cada uma destas palavras-chave servirá como entrada de dados para o modelo proposto, avaliando se os títulos retornados serão os mesmos que foram previamente escolhidos. Para a terceira etapa do modelo foi verificado se a classificação por gênero do site IMDb dos títulos escolhidos serão os mesmos propostos por esta etapa do modelo. Já para a quarta etapa do modelo, foi verificado se o título identificado como semelhante, pela quarta etapa do modelo, aparece também na lista de indicações provida pelo IMDb para o mesmo título. Os nomes dos filmes e séries sofreram alterações para nomes fictícios, substituindo os nomes de todos eles pelos nomes das constelações. O intuito é apenas enfatizar a proposta do modelo e o teor de inovação deste trabalho.

Para a implementação utilizou-se o *Apache Lucene* e *software* OGMA. O *Apache Lucene* é um *software* de busca e contém uma API de indexação de documentos, enquanto o OGMA foi utilizado para identificar os termos relevantes de um documento e também na escolha de títulos semelhantes. As classes foram personalizadas de acordo com o trabalho utilizando a linguagem de programação *Java*. Já o banco de dados foi preparado utilizado o *software opensource* Subliminal. Inúmeras legendas foram coletadas e formatadas de acordo com o que foi exposto na metodologia.

4.1 Benefícios propostos

A criação deste artefato visa demonstrar as diversas utilidades práticas que o modelo pode proporcionar. O modelo facilita e resolve problemas comuns do dia a dia do usuário. Alguns problemas, que são resolvidos com o modelo, são citados abaixo:

- Um usuário procura um filme através de um termo específico, no campo de busca do *Netflix*, esperando encontrar o filme que deseja. O site não retorna nenhum resultado. É necessário o usuário ir até o *Google* para encontrar o nome do filme, para em seguida utilizar novamente o sistema de busca do *Netflix*;
- Um usuário está em uma loja para comprar um filme de presente para um amigo. Sabe de um filme que ele gosta, mas fica na dúvida de qual filme comprar. Pode-se utilizar deste filme preferido para receber uma recomendação;
- Um usuário está em uma locadora precisando alugar um filme que transmita alegria para uma pessoa que por algum motivo está triste. Ele deseja saber quais filmes expressariam melhor aquele sentimento;
- Um usuário deseja comparar a classificação oficial do filme com a de outro sistema;
- Um usuário deseja achar um filme através do nome ou fala de um personagem;
- Um usuário lembra da história do filme, mas não consegue se lembrar do nome do mesmo;

4.2 Busca por palavra-chave

A busca por palavras-chave consiste na pesquisa de um único termo fornecido pelo usuário através da interface do programa. Depois de inserido o termo chave e o usuário clicar em buscar, a classe desenvolvida utilizando o Apache Lucene, retornará como resultado uma lista de filmes e séries que mais representam o termo digitado.

4.2.1 Exemplos práticos

Para o primeiro exemplo, foi realizada a pesquisa com a palavra: amor. O resultado mostra os cinco primeiros títulos ordenados pela classificação, retornando o filme Pavo como o mais relevante. Este filme conta uma história de amor entre uma francesa e um arquiteto japonês que compartilham diferentes perspectivas sobre a guerra. Na sequência aparece também o filme Monoceros, Pisces, um capítulo da série Musca e por último o filme Pictor.

Como segundo exemplo, foi feita uma pesquisa com um personagem de um filme/série. Antes de verificar o resultado abaixo, você saberia dizer os nomes dos filmes que utilizam como nome do personagem o nome “Saul”? Pode ocorrer de você se lembrar do nome de um personagem, mas não do nome do filme ou da série.

O primeiro título citado é o do filme Lyns. Em seguida aparece como resultado quatro episódios do seriado Puppis. É interessante ressaltar que o nome de personagem “Saul” é mais relevante dentro do segundo episódio da terceira temporada, ou seja, provavelmente o nível de interação com este personagem é o maior em comparação com todos os outros da série.

Para o terceiro e último exemplo, foi feita uma pesquisa utilizando uma fala famosa do cinema para verificar o nome do filme ou série que aparecem com este mesmo jargão. Quando queremos relatar para pessoa um filme que vimos, podemos não lembrar o nome, mas sim as cenas, as falas ou mesmo os acontecimentos em geral é que são descritos para que as pessoas possam lembrar se já assistiram ou não aquele título. A frase pesquisada aqui é “Eu sou o rei do mundo”.

Esta frase ficou famosa no filme Fornax, na qual ela aparece no resultado em terceiro lugar. Existem outras duas citações da mesma frase no filme Crater e também no filme Circinus. É importante frisar que apesar da frase ficar reconhecida mundialmente através do ator Leonardo DiCaprio (quando ele abre os braços na frente do navio), o filme em questão é muito extenso, tornando a frase menos relevante que nos outros contextos.

Podemos constatar que são inúmeras as possibilidades, mas cabe ao usuário da ferramenta fazer a pesquisa de acordo com o que necessita e lhe for mais conveniente no momento. Uma diferença sútil na busca pode trazer resultados surpreendentes.

4.2.2 Validação - IMDb

Além dos exemplos citados acima, para validar este tipo de pesquisa optamos por efetuar algumas buscas utilizando as palavras-chave que o site IMDb identifica para cada título.

O primeiro teste foi realizado com o filme Virgo. Escolhemos três palavras-chave que usuários do IMDb utilizam para identificar o filme, são elas: virgo, computação e hacker. Pesquisando cada uma delas.

Para a busca efetuada pelo termo virgo, o resultado mostra como primeira opção o próprio filme Virgo. Em seguida o termo computação traz o filme Virgo na sexta posição. Por último, a busca pelo termo hacker traz o filme Virgo na sétima posição.

O segundo teste de validação foi realizado com o filme Draco. Escolhemos três palavras-chave que o IMDb utiliza para identificar o filme, são elas: dinossauro, ilha e velociraptor.

Para a busca efetuada pelo termo dinossauro, o resultado mostra como primeira opção o próprio filme escolhido. Em seguida o termo ilha traz o filme na sétima posição. Por último, a busca pelo termo velociraptor traz o filme Draco na primeira posição.

O terceiro e último teste foi realizado com o filme Reticulum. As três palavras-chave que mais representam o filme no site do IMDb são: serial killer, FBI (Federal Bureau of Investigation) e psicopata.

Para a busca efetuada pelo termo serial killer o filme em questão não é encontrado. Em seguida o termo FBI traz o filme na sétima posição. Por último, a busca pelo termo psicopata também não apresenta como resultado o nome do filme.

Pode-se concluir que as palavras-chave que os usuários do site IMDb utilizam para identificar o filme Virgo podem também ser utilizadas para efetuar a busca para este projeto proposto. Para os três filmes apresentados, totalizando nove buscas de palavras-chave apenas duas palavras-chave (psicopata e serial killer) não retornaram o nome do filme. Isso pode ser explicado pelo modo como os usuários identificaram o filme, isto é, não pelos acontecimentos, nomes dos personagens ou falas, mas pela percepção que tiveram dos acontecimentos.

4.3 Classificação por gênero

A classificação por gênero, de um título, é um dos fatores que o usuário utiliza para encontrar o filme que deseja. O gênero serve como uma “ponte” para delimitar sua busca dentro de uma categoria específica. É relevante classificar os filmes por gênero devido aos inúmeros títulos disponíveis. A escolha de quais gêneros categorizar foi arbitrária, sendo que nesta seção, foram utilizadas as categorias de Terror e Ação. Quando não for possível classificar em uma destas duas categorias, a classificação genérica padrão “outros” será utilizada. Como adaptação ao modelo, também será apresentado um exemplo referente à análise de sentimentos.

4.3.1 Exemplos e validações com IMDb

Para exemplificar e validar esta etapa do modelo, foram previamente escolhidos quinze filmes, e para cada filme identificado as dez palavras-chave mais relevantes de acordo com o software OGMA. O software OGMA utiliza-se do cálculo TF-IDF, explicado na metodologia, para identificar as palavras-chave mais relevantes. Relembrando, as stop-words são palavras que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido em um contexto pré-definido. Foi necessário também retirar algumas palavras manualmente para que não atrapalhasse o desenvolvimento do modelo proposto, como nome de personagens ou termos variados como “dia”,“noite”,“bem”,“mal”,etc, que não acrescentavam em nada a classificação por gênero.

O IMDb também foi utilizado para demonstrar e validar a categoria de cada filme. Para cada filme representado, nas figuras 20, 21, 22, 23 e 24, foi verificado se o que é proposto pelo modelo vai de acordo com a classificação de gênero identificada pelo IMDb. Após a identificação das palavras-chave pelo OGMA, foi realizado o cruzamento das mesmas com as palavras da “tabela base”. Cada vez que uma palavra-chave era encontrada na “tabela base” o gênero era identificado. No final, para se definir o gênero proposto pelo modelo, era feita uma contagem simples de qual gênero aparecia mais vezes. O de maior número era o gênero escolhido para representar aquele filme específico. Foi demostrado que a maioria dos itens apresentou similaridade com os gêneros definidos pelo IMDb, em contrapartida, ocorreram pequenas discordâncias entre o modelo e o IMDb.

Filme 1	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Andrômeda	Brinquedo Nave Estelar Comando Batata Planeta Cabeça Missão Laser Estranhas	- - - - - - - - - -	Outros	Comédia

Filme 2	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Aguarius	Programa Drogas Laboratório Segurança Diabos Sinal Problema Vivo Alvo Carro	- Ação - Ação Terror - - - Ação Ação	Ação	Ação

Filme 3	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Caelum	Bomba Cidade Homem Polícia Criança Comissário Honra Morrer Morte Dinheiro	Ação - - Ação - - Ação - - Ação	Ação	Ação

Figura 20: Classificação por gênero: Andrômeda, Aquarius e Caelum
Fonte: Elaborada pelo autor

Filme 4	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Crux	Majestade Rei Rainha Diamantes Mosqueteiros Cavalo Matar Guerra Amigo Dama	- - - - - - Ação Ação - -	Ação	Ação

Filme 5	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Gemini	Sargento Vivo Pessoas Torre Mãe Deus Dor Pai Esposa Equipamento	- - - - - Terror Terror - - -	Terror	Drama

Filme 6	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Draco	Parque Dinossauros Sangue Sistema DNA Deus Animais Ilha Segurança Pessoas	- - Terror - - Terror - - Ação -	Terror	Aventura/Ficção

Figura 21: Classificação por gênero: Cruz, Gemini e Draco
Fonte: Elaborada pelo autor

Filme 7	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Columba	Rei Homens Água Majestade Guarda Cavaleiro Garoto Lorde Morte Rochedo	- - - - Ação - - - - -	Ação	Aventura/Drama

Filme 8	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Hyndra	Papai Azul Lua Desastrado Cruel Vila Socorro Gênio Cuidado Essência	- - - - Terror - Terror - - -	Terror	Comédia

Filme 9	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Orion	Aranha Fotografia Cidade Namorada Emprego Trabalho Quadros Ajuda Altura Dinheiro	- - - - - - - - - Ação	Ação	Ação

Figura 22: Classificação por gênero: Columba, Hyndra e Orion
Fonte: Elaborada pelo autor

Filme 10	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Lyra	Assassinato Matar Sistema Chefe Futuro Deus Correr Pessoas Imagens Relatório	Ação Ação - - - Terror - - - -	Ação	Ação

Filme 11	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Cygnus	Família Cão Planeta Capitão Emprego Conselho Experiência Vida Peixe Feliz	- - - - - - - - - -	Outros	Comédia

Filme 12	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Mensa	Arma Matar Mundo Segurança Bomba Sapo Chave Deus Ideia Morte	Ação Ação - Ação Ação - - Terror - Ação	Ação	Ação

Figura 23: Classificação por gênero: Lyra, Cygnus e Mensa
Fonte: Elaborada pelo autor

Filme 13	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Serpens	Máquina Enigma Guerra Deus Trabalho Alemães Mensagens Professor Inteligência Mundo	- - Ação Terror - - - - - -	Outros	Drama

Filme 14	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Octans	Deus Diabo Cristo Espírito Medo Padre Jesus Santo Exorcista Mundo	Terror Terror Terror Terror Terror Terror Terror Terror Terror -	Terror	Terror

Filme 15	Palavras-chave	Relação com a tabela base	Gênero (modelo)	Gênero (IMDB)
Lepus	Namorados Amor Feliz Flores Problema Ouça Coração Namorado Amigo Especial	- - - - - - - - - -	Outros	Romance

Figura 24: Classificação por gênero: Serpens, Octans e Lepus
Fonte: Elaborada pelo autor

A figura 25 expressa bem os resultados encontrados. O nível de aceitação da categorização dos títulos por legendas foi alto. De cento e cinquenta palavras-chave analisadas, quarenta palavras-chave foram relacionadas. Dos quinze filmes analisados, onze apresentaram o mesmo gênero e em quatro os resultados foram divergentes. Vale ressaltar que os títulos avaliados são os das categorias terror e ação. Os títulos que não se encaixavam nestes gêneros foram categorizados como outros, confirmando que o modelo demonstrou que estes títulos não fazem parte dos gêneros terror e ação. Isso gerou uma taxa de similaridade de classificação por Gênero de 73%. A palavra mais encontrada dentro dos filmes foi a palavra "Deus", provavelmente porque pode ser usada em vários sentidos dentro dos títulos.

Resumo	
Palavra-chave mais encontrada	"Deus"
Total de Palavras-Chave analisadas	150
Porcentagem de palavras-chave relacionadas	40
Porcentagem de filmes classificados iguais ao IMDB	73%

Figura 25: Compilação dos resultados - segunda etapa
Fonte: Elaborada pelo autor

4.3.2 Análise de sentimentos

Para exemplificar a análise de sentimentos, foi utilizada a mesma classe Java de pesquisa da primeira etapa do modelo. As palavras-chave da "tabela base" (apêndice D) de cada sentimento foram utilizadas para efetuar a busca. O filme ou série que apresentar a maior média de pontuação, entre os termos pesquisados para um sentimento, será o que melhor representará cada sentimento. É importante ressaltar que esta pesquisa foi realizada no banco de dados com mais de 700 arquivos de legendas entre filmes e séries.

O resultado para cada sentimento é demonstrado na figura 26. A análise de sentimentos representa um caminho bastante promissor. Através do exemplo prático, utilizando a "tabela base" criada, o resultado se mostrou bastante coerente. A série *Musca* expressa justamente tristeza pelos casos raros de doenças de seus pacientes e o sentimento de vergonha

por causa dos erros em alguns diagnostico médicos. *Cetus* é um filme de comédia, representando assertivamente o sentimento de felicidade. A série *Columba* representa uma época de conflitos entre vários reinos ficando classificado com o sentimento de "Irritado". Ainda temos como destaque o filme *Taurus* que expressa o sentimento de "Encantado". Já os filmes *Tucana* e o seriado *Scutum* foram classificados como "Satisfeito" e "Animado" respectivamente. Todos estes resultados demonstram na prática o poder da análise de sentimentos, podendo surgir diversas ramificações de modelo para as mais diversas áreas.

Compilação dos resultados	
Sentimentos	***Títulos recomendados***
Triste	Musca
Feliz	Cetus
Irritado	Columba
Tenso	Sagitta
Angustiado	Vela
Entendiado	Dorado
Cansado	Auriga
Sonolento	Musca
Sereno	Columba
Satisfeito	Tucana
Encantado	Taurus
Animado	Scutum

Figura 26: Compilação dos resultados - análise de sentimentos
Fonte: Elaborada pelo autor

4.4 Títulos semelhantes

Umas das maiores dificuldades é a de identificar rapidamente títulos que sejam do interesse do usuário sem que este realize um grande esforço, ou ainda, que seja necessária a indicação de um amigo. Os inúmeros filmes disponíveis de diversas categorias deixam perdido qualquer pessoa que deseja assistir um filme. Não necessariamente apenas usuários podem aproveitar deste modelo, mas empresas que desejam construir um catálogo mais organizado podem criar um novo tipo de organização que identifique mais sentido nas relações entre os filmes e séries.

Esta proposta tenta ajudar ao usuário a encontrar automaticamente um filme com um enredo semelhante, ou próximo, a um tema que ele deseja. Serão analisados os resultados dos testes para verificar se o modelo foi condizente com o tipo de filme.

4.4.1 Exemplos e validações com o IMDb

Foram utilizados três filmes para demonstrar a terceira e última etapa do trabalho. Estes filmes foram analisados com o OGMA trazendo como resultado os 5 termos mais relevantes. Cada termo foi pesquisado dentro da base de dados deste trabalho para que retornasse os títulos mais relevantes. No final o filme mais citado nos resultados era o filme indicado.

Para validar a indicação realizada por este modelo, optamos por comparar os filmes que os próprios usuários do IMDb indicam. Com a mensagem no site de que “pessoas que gostaram deste filme gostaram também de” os usuários conseguem indicações para que possam assistir outros filmes. No site constam até 12 indicações para cada filme.

O primeiro filme escolhido será o Chamaeleon, conforme é demonstrado na figura 27. Para este filme o IMDb apresenta indicações como: Câncer, Cepheus e a Carina. O resultado do modelo não é condizente com nenhuma indicação do site do IMDb. Os dois filmes apresentam um cenário bem distinto apesar de terem um diálogo similar, tendo principalmente as palavras “guerra” e “máquina” em seus contextos.

Filme: Chamaeleon		
Termo	**Filmes encontrados**	**Resultado**
Máquina	Eridanus	0.7132805
	Antlia	0.1634331
	Apus	0.12858847
	Aquila	0.118284196
	Virgo	0.1008731
Enigma	Musca	0.26404357
	Ara	0.12447132
	Canis Major	0.10779533
	Cassiopeia	0.10163041
	Coma Berenices	0.089829445
Guerra	Eridanus	0.59853584
	Equuleus	0.29926792
	Horologium	0.2591736
	Leo Minor	0.1692915
	Lupus	0.14963396
Deus	Microscopium	0.15179044
	Norma	0.13076155
	Ophiuchus	0.12930048
	Pegasus	0.12134158
	Phoenix	0.10896795
Trabalho	Piscis Austrinus	0.17129488
	Triangulum Australe	0.12818536
	Sextans	0.11330091
	Telescopium	0.10901139
	Vulpecula	0.10594004
Filme indicado / Total	**Eridanus**	**1,31181634**

Figura 27: Título escolhido: Chamaeleon - Recomendação: Eridanus
Fonte: Elaborada pelo autor

O segundo filme escolhido será o *Reticulum*, conforme é demonstrado na figura 28. Para este filme o IMDb apresenta indicações como: *Fornax*, *Lepus* e o título *Ophiuchus*. O resultado do modelo não é condizente com nenhuma indicação do site do IMDb. Os dois filmes são para públicos diferentes. O primeiro é ambientado no mundo dos vampiros, mas é voltado para o público infantil, já o segundo é um filme de ação e terror, sendo indicado para adultos. As palavras “drácula” e “monstros” são as que mais se relacionam e aparecem em seus contextos.

Filme: Reticulum		
Termo	**Filmes encontrados**	**Resultado**
Drácula	Sculptor	0.31268686
	Camelopardalis	0.12765387
	Capricornus	0.12765387
	Boötes	0.10637823
	-	-
	-	-
Humano	Cassiopeia	0.15102927
	Canes Venatici	0.1483078
	Corona Australis	0.1353859
	Hydrus	0.12863639
	Lacerta	0.1248706
Festa	Monoceros	0.2410086
	Leo	0.20871955
	Ophiuchus	0.18075645
	Sextans	0.16827509
	Piscis Austrinus	0.16755442
Monstros	Hydra	0.1997026
	Volans	0.1997026
	Ursa Minor	0.17861944
	Sculptor	0.15850902
	Triangulum	0.13209085
Hotel	Aries	0.20439863
	Corona Borealis	0.17006993
	Corvus	0.15288898
	Canes Venatici	0.1344521
	Circinus	0.12676266
Filme indicado / Total	**Sculptor**	**0,47119588**

Figura 28: Título escolhido: Reticulum- Recomendação: Sculptor
Fonte: Elaborada pelo autor

O terceiro filme escolhido será o *Auriga*, conforme é demonstrado na figura 29. Para este filme o IMDb apresenta indicações como: *Monoceros, Piscis Austrinus* e o título *Taurus*. O resultado do modelo não é condizente com nenhuma indicação do site do IMDb. Os dois filmes são de histórias fictícias e possuem cenários parecidos. No IMDb não indicam séries a partir dos filmes, então pode ser esta a justificativa para não aparecer na indicação.

Filme: Auriga		
Termo	**Filmes encontrados**	**Resultado**
Varinha	Ophiuchus	0.25490347
	Lacerta	0.19744737
	Musca	0.12091132
	Cetus	0.11399629
	Horologium	0.07053161
Espada	Ursa Minor	0.27417326
	Sextans	0.24460164
	Tucana	0.23500565
	Columba	0.158294
	Caelum	0.1599011
Castelo	Columba	0.31965613
	Lynx	0.22029704
	Columba	0.21856919
	Sculptor	0.21770014
	Grus	0.19374087
Professor	Apus	0.23003149
	Aquila	0.1840252
	Eridanus	0.18310277
	Delphinus	0.17458165
	Cepheus	0.16587807
morrer	Scutum	0.11702266
	Musca	0.10910869
	Musca	0.10585098
	Scutum	0.09979726
	Corona Australis	0.095412776
Filme indicado / Total	**Columba**	**0,69651932**

Figura 29: Título escolhido: Auriga - Recomendação: Columba
Fonte: Elaborada pelo autor

Pode-se concluir que a identificação de títulos semelhantes não foi tão efetiva quanto os dois modelos propostos anteriormente. Este modelo consegue identificar apenas os termos utilizados nos diálogos de cada título, mas não consegue demonstrar uma visão mais abrangente sobre cada filme para poder indicar novos filmes. São várias as razões que podem explicar o ocorrido como:

- duplo sentido das palavras, onde cada uma estará em um contexto ou situação diferente;
- uma mesma pessoa pode gostar de filmes completamente diferentes;
- uma mesma pessoa pode gostar de um filme em um dado momento, mas nem tanto em outro momento;
- as pessoas são distintas, o que é bom para uma pode não ser para a outra;
- existem filmes com poucos diálogos, dificultando a precisão do modelo;
- a base de dados dos títulos deve ser atualizada constantemente;

A figura 30 resume o resultado desta etapa.

Filmes	Palavras-chave	Título Indicado	Coincide com indicações do IMDB?
Chamaeleon	Máquina Enigma Guerra Deus Trabalho	Eridanus	Não
Reticulum	Drácula Humano Festa Monstros Hotel	Sculptor	Não
Auriga	Varinha Espada Castelo Professor Morrer	Columba	Não

Figura 30: Compilação dos resultados - terceira Etapa
Fonte: Elaborada pelo autor

5 CONSIDERAÇÕES FINAIS

Hoje o acervo de dados e informações é imenso, bem como os diversos tipos de buscas que podem ser realizadas. É preciso evoluir no sentido do refinamento dos dados que os bancos de dados disponibilizam, para que os usuários possam obter resultados mais precisos. Como consequência poderemos ter novas ferramentas de recomendação muito mais assertivas. Foi proposto neste trabalho um modelo de recomendação de conteúdo com novas possibilidades de recuperação da informação utilizando como base de dados o conteúdo de arquivos de legendas de filmes e séries.

É preciso achar alternativas em meio a tanta informação disponível, sendo uma destas alternativas a criação de sistemas de recomendação de conteúdo para ajudar e auxiliar aos usuários. Esta dissertação espera como resultado demonstrar que recomendações por conteúdo podem ser muito mais eficazes e relevantes do que apenas os modelos parametrizados de padrões de busca tradicionais. É possível muito mais, proporcionando novos modelos de ferramentas aos usuários possibilitando alternativas de pesquisa, classificação e recomendação de documentos.

O modelo foi apresentado em quatro etapas distintas: os passos iniciais que compõe os requisitos para criação do modelo, um mecanismo de busca utilizando palavras-chave, a classificação de filmes e séries por gênero e a identificação de títulos similares. A segunda e quarta etapa recomendam diretamente títulos para o usuário. A terceira etapa (classificação por gênero) facilita a busca do usuário atuando como fator inicial no processo de seleção de um filme ou série. Todas as etapas utilizaram um único banco de dados, provido de informação das próprias falas dos personagens, mais precisamente dos arquivos de legendas, aproximando os usuários dos filmes e séries que mais lhe interessam e possam lhe agradar naquele momento.

Para segunda etapa do modelo, concluiu-se que os resultados foram significativos já que proporcionam ao usuário liberdade de pesquisa dentro de um conteúdo específico. São variadas as formas que se pode pensar para realizar a busca trazendo resultados surpreendentes.

Já na terceira etapa do modelo foi apresentada a classificação automática de gênero onde a mesma apresentou um índice de acerto de 73%. Este índice é importante, visto que, esta foi apenas a primeira implementação do modelo, podendo ocorrer inúmeras melhorias cada vez que é testado. Outro destaque importante é a “tabela base” de gêneros. Esta tabela foi criada

como referência para a classificação por gênero, mas foi bastante aprimorada no decorrer dos testes. Isso resultou em uma classificação de gênero mais assertiva em relação ao site IMDb.

Por último, a quarta etapa recomendou títulos similares através de um título previamente escolhido. Foi criado e modelado um tipo de abordagem em cima dos arquivos de legendas, além da segunda etapa do modelo ser também utilizada neste processo. Os resultados foram abaixo do esperado, retratando a dificuldade para se criar um software de recomendação de filmes. A principal dificuldade desta etapa foi em recomendar títulos de temas alternativos ao título pré-escolhido e que possam representar o real gosto do usuário. Esta etapa do modelo demonstrou apenas resultados padrões, isto é, os títulos recomendados contêm praticamente o mesmo cenário e temática.

Com o modelo, verificaram-se as vantagens de recomendar títulos que não sejam da forma convencional, pelo nome do autor, diretor, nome do filme, entre outros, mas sim pelo conteúdo em si das falas dos personagens. A ideia central do trabalho foi atingida com o objetivo de abrir novos leques de estudo dentro da recuperação da informação em conjunto com a recomendação da informação baseado em conteúdos específicos. O trabalho é baseado nos arquivos de legendas, sem qualquer outra fonte. Para um sistema de recomendação robusto, este modelo deve ser empregado em conjunto com outras fontes e etapas de trabalho. É o começo de um grande algoritmo.

Vale ressaltar o teor inovador deste trabalho que conta com a análise de texto de arquivos e legendas para recuperar informação. A ação de criar um banco de dados de arquivos de legendas é promissora já que a pesquisa pode ser realizada em cima de conteúdo e não apenas de dados parametrizados. Desta forma as possibilidades que poderão ser ofertadas aos usuários são gigantescas. Uma destas possibilidades foi demonstrada neste trabalho com a recomendação de conteúdo e com a adaptação do modelo para a análise de sentimentos. Na análise de sentimentos foi possível inferir alguns sentimentos de acordo com as palavras-chave pré-estabelecidas. A análise de sentimentos demonstrou recomendações com coesão, determinando, na maioria das vezes, títulos apropriados para cada sentimento. Trata-se de uma possibilidade inovadora com um longo caminho a ser percorrido.

O estudo realizado apresentou algumas limitações. Primeiro quanto à coleta de dados. A coleta de dados foi realizada utilizando um *software* baseado em duas listas de arquivos textos pré-definidos, dando preferências a títulos mais bem ranqueados dentro do site IMDb. Apesar disso, inúmeros filmes e séries ficaram de fora da base de dados, seja porque

não entraram nesta pré-lista ou ainda porque o *software* utilizado para download das legendas não continha em sua base de dados o título indicado.

Outra limitação no trabalho foi na identificação das palavras-chave para cada título. Mesmo com o filtro realizado pelo software OGMA (retirando as *stop-words*), inúmeros termos soltos aparecem como mais relevante na análise dos arquivos de legendas. Desta forma, foi necessário um filtro manual para cada título analisado para conseguirmos identificar as palavras-chave mais relevantes.

Como última limitação temos a construção das "tabelas base". A construção das tabelas foi realizada dentro do processo de identificação das palavras-chave nos arquivos de legendas. Para complementá-las foram realizadas pesquisas na Internet de termos para categorizar a identificação dos gêneros e sentimentos. Também foram identificados alguns trabalhos que ajudaram na formulação das "tabelas base", mas denota-se que todo o processo foi realizado de forma manual, sem uma fórmula padrão. Foi este conjunto de tentativas composto por diferentes métodos que se chegou à montagem final das "tabelas base".

Para trabalhos futuros, é recomenda-se a implementação de um algoritmo que contemple todas as fases do modelo de forma automática, retirando as análises manuais. Para a coleta de dados recomenda-se pesquisar uma alternativa para que o banco de dados seja sempre atualizado e não estático. É interessante também mudar o número de palavras-chave analisadas para verificar se haverá alteração nos resultados. Neste trabalho os testes foram realizados com dez (classificação por gênero) e cinco palavras-chave (similaridade) respectivamente.

Outra recomendação de trabalho futuro é a de possibilitar ao usuário a liberdade de escolha da língua que se quer utilizar para pesquisa nos arquivos de legenda: inglês, espanhol, entre outras. Atendendo a este e aos outros requisitos o sistema poderá ser utilizado automaticamente em larga escala atendendo inúmeros usuários ao redor do mundo.

Finalizando, é indicada a criação de um modelo robusto que contemple a área de análise de sentimentos, já que neste trabalho foi apenas demonstrado a possibilidade, podendo ainda ser muito mais explorada. A possibilidade de extrair de um filme ou série quais os sentimentos que eles despertam pode ser bastante promissor.

REFERÊNCIAS

ADHIKARI, V. K. *et al*. ***Unreeling netflix****: Understanding and improving multi-cdn movie delivery*. *In*: IEEE. INFOCOM, 2012 Proceedings IEEE. [S.l.], 2012. p. 1620–1628.

AKEN, J. E. V.; ROMME, G. ***Reinventing the future****: adding design science to the repertoire of organization and management studies.* Organization Management Journal,Taylor & Francis, v. 6, n. 1, p. 5–12, 2009.

ALAN, R. H. von *et al*. ***Design science in information systems research***. MIS quarterly, Springer, v. 28, n. 1, p. 75–105, 2004.

ANSARI, A.; ESSEGAIER, S.; KOHLI, R. ***Internet recommendation systems***. p. 363–375,2000.

ARAUJO, G. D. de *et al*. **Análise de sentimentos sobre temas de saúde em mídia social**. Journal of Health Informatics, v. 4, n. 3, 2012.

BAEZA-YATES, R.; RIBEIRO-NETO, B. ***Modern information retrieval***. ACM press, 1999. 513 p. ISSN 0022541X. ISBN 020139829X. Disponível em: <http: //web.simmons.edu/~benoit/LIS466/Baeza-Yateschap01.pdf$\delimiter"026E30F$nftp: //mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and\ _visualization-modern_ir.pdf>.

BALABANOVIc, M.; SHOHAM, Y. Fab: ***Content-based, collaborative recommendation. Commun***. ACM, ACM, New York, NY, USA, v. 40, n. 3, p. 66–72, mar. 1997. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/245108.245124>.

BARION, E. C. N.; LAGO, D. **Mineração de textos**. Revista de Ciências Exatas e Tecnologia, v. 3, n. 3, p. 123–140, 2015.

BARREIRO, S. C**. Sonar, sdi automatizado do centro de informações nucleares**. Revista de Biblioteconomia de Brasília, Brasília, v. 6, n. 2, p. 139–154, 1978.

BARRETO, J. S. **Anotação automática e recomendação personalizada de documentários brasileiros-sistema docunb**. 2011.

BARRETO, J. serra. **Desafios e avanços na recuperação automática da informação audiovisua**l. Ci. Inf, SciELO Brasil, v. 36, n. 3, p. 17–28, 2007.

BARTH, I. J. **Modelando o perfil do usuário para a construção de sistemas de recomendação**: um estudo teórico e estado da arte. Revista de Sistemas de Informação da FSMA, v. 6, p. 59–71, 2010.

BECKER, K.; TUMITAN, D. **Introdução à mineração de opiniões**: Conceitos, aplicações e desafios. Simpósio Brasileiro de Banco de Dados, 2013.

BRASCHER, M. **A ambiqüidade na recuperação da informação**. IASI, 2002.

BROWN, A.; FUGMANN, R.; SUENONIUS, E. UNISIST DRAFT ON INDEXING PRINCIPLES-TEXT AND COMMENTS. [S.l.]: INT SOC KNOWLEDGE ORGANIZATION 36A WOOGSTR, D-6000 FRANKFURT 50, GERMANY, 1977. 29–34 p.

CAMPOS, L. F. d. B. **Web 2.0, biblioteca 2.0 e ciência da informação: um protótipo para disseminação seletiva de informação na web utilizando mashups e feeds rss**. 2012.

CARDOSO, O. N. P. **Recuperação de informaçã**o. INFOCOMP: Journal of Computer Science, v. 2, n. 1, 2000.

CARDOSO, O. N. P. **Recuperação de informação**. INFOCOMP Journal of Computer Science, v. 2, n. 1, p. 33–38, 2004.

CARITÁ, E. C. *et al*. **Implementação e avaliação de um sistema de gerenciamento de imagens médicas com suporte à recuperação baseada em conteúdo**. Radiol Bras, SciELO Brasil, v. 41, n. 5, p. 331–6, 2008.

CAZELLA, S. C.; NUNES, M.; REATEGUI, E. B**. A ciência da opinião**: Estado da arte em sistemas de recomendação. In: XXX Congresso da SBC Jornada de Atualização da Informática. [S.l.: s.n.], 2010.

CHELLA, M. T. **Sistema para classificação e recuperação de conteúdo multimídia baseado no padrão mpeg-7**. UNICAMP: São Paulo, v. 10, 2004. Citado na página 46.
CHEN, H. Semantic research for digital libraries. D-Lib Magazine, 1999.

CHOO, C. W.; ROCHA, E. **A organização do conhecimento**: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões São Paulo: SENAC, 2003.

CHOWDHURY, G. ***Introduction to modern information retrieva***l. [S.l.]: Facet Publishing,2004. 496 p.

COADIC, Y.-F. L.; GOMES, M. Y. F. **A ciência da informação**. [S.l.]: Briquet de lemos Livros, 1996.

CORUMBA, D.; MACEDO, H. **Categorização automática de mensagens de call-for-papers**.Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.5329/RESI, v. 10, n. 2, 2011.

DATAXIS: Site. 2015. Disponível em: <http://dataxis.com/mercado-ott-en-america-latina-2013-2018/>. Acesso em: 19 agosto 2015.

DIMARTINO, D.; ZOE, L. R. ***End-user full-text searching***: Access or excess? Library & information science research, Elsevier, v. 18, n. 2, p. 133–149, 1996.

DOSZKOCS, T. E.; REGGIA, J.; LIN, X. ***Connectionist models and information retrieval***. Annual review of information science and technology, Information Today, v. 25, p. 209–262, 1990.

FERNANDES, R. P. M. et al. **Panorama atual do uso dos mecanismos de busca na web**. 2013.

FERNEDA, E. **Recuperação de informação: análise sobre a contribuição da ciência da computação para a ciência da informação**. Tese (Doutorado) — Escola de Comunicações e Artes da Universidade de São Paulo, 2003.

FERNEDA, E. **Aplicando algoritmos genéticos na recuperação de informação**. Revista de Ciência da Informaçao, v. 10, 2009.

FERREIRA, E. d. B. A. **Análise de sentimento em redes sociais utilizando influência das palavras**. Trabalho de Graduação-Universidade Federal de Pernambuco-UFPE. Departamento de Ciência da Computação, v. 22, p. 23–25, 2010.

FOSKETT, A. C. ***Subject approach to information***. Conn.], Linnet Books, 1972.

GEY, F. F. ***Models in information retrieval.*** Folders of Tutorial Presented at the 19th ACM Conference on Research and Development in Information Retrieval1992. Folder.

GÓMEZ, A. B. *et al*. ***An efficient and scalable recommender system for the smart web***. IEEE. Computer Society, 2015.

GOULARTE, R**. Personalização e adaptação de conteúdo baseadas em contexto para TV Interativa**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, 2003.

GUERRA, P. H. C. et al. **From bias to opinion**: a transfer-learning approach to real-time sentiment analysis. *In*: ACM. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. [S.l.], 2011. p. 150–158.

HALLINAN, B.; STRIPHAS, T. ***Recommended for you****: The netflix prize and the production of algorithmic culture. new media & society*, SAGE Publications, v. 18, n. 1, p. 117–137, 2016.

HAN, J.; KAMBER, M.; PEI, J. ***Data mining****: concepts and techniques*Elsevier, 2011.

HERLOCKER, J. L.; KONSTAN, J. A.; RIEDL, J. ***Explaining collaborative filtering recommendations***. *In*: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. New York, NY, USA: ACM, 2000. (CSCW '00), p. 241–250. ISBN 1-58113-222-0. Disponível em: <http://doi.acm.org/10.1145/358916.358995>.

IBM: Site. 2015. Disponível em: <http://www.ibm.com/developerworks/br/java/library/ os-apache-lucenesearch/>. Acesso em: 26 jan. 2015.

IMDB: Site. 2015. Disponível em: <http://www.imdb.com/>. Acesso em: 18 novembro

2015.

KARMAKER, D. *et al.* **An automated music selector derived from weather condition and its impact on human psychology**. GSTF Journal on Computing (JoC), Global Science and Technology Forum, v. 4, n. 3, p. 13, 2015.

KOREN, J.; ZHANG, Y.; LIU, X. **Personalized interactive faceted search**. *In*: ACM. Proceedings of the 17th international conference on World Wide Web. [S.l.], 2008. p. 477–486.

KURAMOTO, H. **Sintagmas nominais**: uma nova proposta para a recuperação de informação. IASI, 2002.

LANCASTER, F. W. ***Information retrieval systems***. 1968.

LIU, B. ***Sentiment analysis and subjectivity***. Handbook of natural language processing, v. 2, p. 627–666, 2010.

LOEB, S.; TERRY, D. ***Information filtering***. Communications of the ACM, ACM, v. 35, n. 12, p. 26–28, 1992.

LUHN, H. P**. A business intelligence system**. IBM Journal of Research and Development,v. 2, 1958.

MAIA, L. C. G.; SOUZA, R. R. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. Perspectivas em Ciência da Informação, v. 15, n. 1, p. 154–172, 2010.

MAIA, L. C. G.; SOUZA, R. R. Medidas de similaridade em documentos eletrônicos. 2013.

MARQUES, T. M. Abordagens de recomendação para a recuperação de perfis: uma proposta de modelo. 2007.

MATTMANN, C. A. et al. Software connector classification and selection for data-intensive systems. In: IEEE COMPUTER SOCIETY. Proceedings of the Second International Workshop on Incorporating COTS Software into Software Systems: Tools and Techniques. [S.l.], 2007. p. 4.

MCDONALD, D. W. Recommending collaboration with social networks: A comparative evaluation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 2003. (CHI '03), p. 593–600. ISBN 1-58113-630-7. Disponível em: <http://doi.acm.org/10.1145/642611.642714>.

MELVILLE, P.; GRYC, W.; LAWRENCE, R. D. Sentiment analysis of blogs by combining lexical knowledge with text classification. In: ACM. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. [S.l.], 2009. p. 1275–1284.

MOHAMMAD, S. M.; TURNEY, P. D. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. [S.l.], 2010. p. 26–34.

MONTANER, M.; LÓPEZ, B.; ROSA, J. L. D. L. A taxonomy of recommender agents on the internet. Artificial intelligence review, Springer, v. 19, n. 4, p. 285–330, 2003.

MOOERS, C. N. Zatocoding applied to mechanical organization of knowledge. American documentation, Wiley Online Library, v. 2, n. 1, p. 20–32, 1951. Citado na página 11.

MOURA, M. A. Folksonomias, Redes Sociais e a Formação par o Tagging Literacy: Desafios para a Organização da Informação em Ambientes Colaborativos Virtuais. Informação & Informação, v. 14, p. 25–45, 2009. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/2196>.

NA, J.-C. et al. A sentiment-based digital library of movie review documents using fedora/une bibliothèque numérique de documents critiques de films basée sur les sentiments en utilisant fedora. Canadian Journal of Information and Library Science, University of Toronto Press, v. 35, n. 3, p. 307–337, 2011.

NETFLIX: Site. 2015. Disponível em: <http://brasilblog.netflix.com/2014/09/ uma-nova-experiencia-de-busca-no-site.html>. Acesso em: 09 junho 2015.

NODARI, A. R. Os sistemas de recomendação como instrumento para atingir mercados de nicho. 2014.

OGMA: Site. 2016. Disponível em: <http://www.luizmaia.com.br/ogma/>. Acesso em: 15 março 2016.

OLETO, R. R. et al. Percepção da qualidade da informação. Ciência da informação, SciELO Brasil, v. 35, n. 1, p. 57–62, 2006.

OSIEK, B. A. RECONHECIMENTO DE SENTIMENTO EM TEXTO ABORDADO ATRAVÉS DA COMPUTAÇÃO AFETIVA. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2014.

PERES, S. M.; BOSCARIOLI, C. Sistemas gerenciadores de banco de dados relacionais fuzzy: uma aplicação em recuperação de informação. Acta Scientiarum. Technology, v. 24, p. 1733–1743, 2008.

PERUGINI, S.; GONÇALVES, M. A.; FOX, E. A. Recommender systems research: A connection-centric survey. J. Intell. Inf. Syst., Kluwer Academic Publishers, Hingham, MA, USA, v. 23, n. 2, p. 107–143, set. 2004. ISSN 0925-9902. Disponível em: <http://dx.doi.org/10.1023/B:JIIS.0000039532.05533.99>.

PINHEIRO, L. V. R.; LOUREIRO, J. M. M. Traçados e limites da ciência da informação. Ciência da Informação, 1995.

REATEGUI, E. B.; CAZELLA, S. C. Sistemas de recomendaçao. In: CITESEER. XXV Congresso da Sociedade Brasileira de Computação. Universidade do Vale do Rio dos Sinos (UNISINOS). São Leopoldo. [S.l.], 2005.

REIS, J. C. et al. Uma abordagem multilıngue para análise de sentimentos. 2012.

REIS, U. S. MOFI: Modelo Computacional para Recuperação de Informação baseado em Ontologias, Folksonomia e Indexação automatica de conteúdo. 136 p. Tese (Doutorado) — SENAI CIMATEC, 2011. Citado na página 22.

REIS, U. S.; PEREIRA, H. B. B. Mofi: Um modelo para recuperação de informação baseado em ontologias, folksonomia e indexação automática de conteúdo. Seminário de Pesquisa em Ontologia no Brasil, 2010. Citado na página 22.

RESNICK, P.; VARIAN, H. R. Recommender systems. Communications of the ACM, ACM, v. 40, n. 3, p. 56–58, 1997. Citado na página 35.

ROBREDO, J. A indexação automática de textos: o presente já entrou no futuro. Estudos Avançados em Biblioteconomia e Ciência da Informação. Brasília, ABDF, v. 1, p. 236–74, 1982.

RODRIGUES, B. C.; CRIPPA, G. et al. A recuperação da informação e o conceito de informação: o que é relevante em mediação cultural? Perspectivas em Ciência da Informação, SciELO Brasil, v. 16, n. 1, p. 45–64, 2011.

SALTON, G.; MCGILL, M. J. Introduction to modern information retrieval. Introduction to modern information retrieval, 1983. ISSN 0070544840. Disponível em: <http://search. ebscohost.com/login.aspx?direct=true&db=lxh&AN=ISTA2001897&site=ehost-live>.

SANTOS, L. M. et al. Twitter, análise de sentimento e desenvolvimento de produtos: Quanto os usuários estão expressando suas opiniões? Revista PRISMA. COM, n. 13, 2010.

SANT'ANA, R. A importância do papel do profissional da ciência da informação nos processos de recuperação de conteúdos digitais estruturados. Ensino e pesquisa em biblioteconomia no Brasil: a emergência de um novo olhar. Marília: Cultura acadêmica, p. 145–154, 2008.

SARACEVIC, T. Information Science. JASIS – Journal of the American Society for Information Science, v. 50, n. 12, p. 1051-1063, 1999.

SCHAFER, J. B.; KONSTAN, J.; RIEDI, J. Recommender systems in e-commerce. Proceedings of the 1st ACM conference on Electronic commerce EC 99, v. 2001, p. 158–166, 1999. Disponível em: <http://portal.acm.org/citation.cfm?doid=336992.337035>.

SHAW, I. S. Controle e modelagem fuzzy. [S.l.]: Edgard Blucher, 1999.

SILVA, F. S. da; ALVES, L. G. P.; BRESSAN, G. Personaltvware: Uma proposta de arquitetura sensível ao contexto para suporte a recomendação personalizada de conteúdo

no cenário da tv digital interativa (position paper). SBCUP–I Simpósio Brasileiro de Computação Ubíqua e Pervasiva, Bento Gonçalves, Rio Grande do Sul, 2009.

SILVA, M. d. R. da; FUJITA, M. S. L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. Transinformação, v. 16, n. 2, 2012.

SILVA, M. P. d. S. Mineração de dados: Conceitos, aplicações e experimentos com weka. Sociedade Brasileira de Computação, v. 1, 2004.

SILVA, N. J. S. de Almeida Neves da. Descoberta automÁtica de temas utilizando legendas. 2012.

SJÖBERG, M. et al. The mediaeval 2014 affect task: Violent scenes detection. In: MediaEval. [S.l.: s.n.], 2014.

SMITH, N. A.; BAMMAN, D.; OCONNOR, B. Learning latent personas of film characters. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2013.

SORDI, J. O. D.; AZEVEDO, M. C. de; MEIRELES, M. A pesquisa design science no brasil segundo as publicações em administração da informação. Revista de Gestão da Tecnologia e Sistemas de Informação, SciELO Brasil, v. 12, n. 1, p. 165–186, 2015.

SOUTO, L. F. Disseminação seletiva de informações: discussão de modelos eletrônicos. Dissertação (Mestrado) — Pontifícia Universidade Católica de Campinas, 2003.

SOUTO, L. F. Disseminação seletiva de informações: discussão de modelos eletrônicos. 2012.

SOUZA, M. V. d. S. Mineração de opiniões aplicada a mídias sociais. Pontifícia Universidade Católica do Rio Grande do Sul, 2012.

SOUZA, R. R.; ALVARENGA, L. A web semântica e suas contribuições para a ciência da informação. Ciência da Informação, Brasília, SciELO Brasil, v. 33, n. 1, p. 132–141, 2004.

SOUZA, R. R. et al. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. Perspectivas em ciência da informação, SciELO Brasil, v. 11, n. 2, 2006.

SPOTIFY: Site. 2015. Disponível em: <https://news.spotify.com/br/>. Acesso em: 19 agosto 2015.

TEIXEIRA, J. d. F. Mentes e máquinas: uma introdução à ciência cognitiva. Porto Alegre: Artmed, 1998.

TERVEEN, L. et al. Phoaks: A system for sharing recommendations. Communications of the ACM, ACM, v. 40, n. 3, p. 59–62, 1997.

THEGUARDIAN: Site. 2015. Disponível em: <http://www.theguardian.com/film/2014/

jun/04/film-streaming-downloads-dvd-netflix>. Acesso em: 26 jan. 2015.

VELSEN, L. van; MELENHORST, M. Incorporating user motivations to design for vídeo tagging. Interacting with Computers, Oxford University Press, v. 21, n. 3, p. 221–232, 2009.

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), 1988.

YAHIAOUI, I. Construction automatique de résumés vidéos: proposition d'une méthode générique d'évaluation. 2003.

APÊNDICE A – Stop-works

ainda	ela	melhor	após	são	ninguém	todos	tinha	desta	foi	poderiam	desta	mais	porque
alem	elas	bom	aquela	se	no	tua	tinham	deste	for	podia	destas	mas	portanto
ambas	ele	claro	aquelas	seja	nos	tuas	toda	dispoe	foram	podiam	deste	mediante	proprio
ambos	eles	preciso	aquele	sejam	nós	tudo	todas	dispoem	fosse	pois	deste	menos	propios
antes	em	algo	aqueles	sem	nossa	um	todavia	diversa	fossem	por	destes	mesma	quais
ao	enquanto	podemos	aquilo	sempre	nossas	uma	todo	diversas	grande	porém	deve	mesmas	qual
aonde	entre	comigo	as	sendo	nosso	umas	todos	diversos	grandes	porque	devem	mesmo	qualquer
aos	era	quê	até	será	nossos	uns	tu	do	há	posso	devendo	mesmos	quando
apos	essa	tão	através	serão	num	a	tua	dos	isso	pouca	dever	na	quanto
aquele	essas	doids	cada	seu	numa	à	tuas	durante	isto	poucas	deverá	nas	que
aqueles	esse	aí	coisa	seus	nunca	agora	tudo	e	já	pouco	deverão	nao	quem
as	esses	às	coisas	si	o	ainda	última	ela	la	poucos	deveria	nas	quer
assim	esta	los	com	sido	os	alguém	últimas	elas	la	primeiro	deveriam	nem	se
com	está	i	como	só	ou	algum	último	ele	lá	primeiros	devia	nesse	seja
como	estamos	cá	contra	sob	outra	alguma	últimos	eles	lhe	própria	deviam	neste	sem
contra	estão	pela	contudo	sobre	outras	algumas	um	em	lhes	próprias	disse	nos	sendo
contudo	estas	pelas	da	sua	outro	alguns	uma	entao	lo	próprio	disso	o	seu
cuja	estava	pelo	daquele	suas	outros	ampla	umas	entre	mas	próprios	disto	os	seus
cujas	estavam	pelos	daqueles	talvez	para	amplas	uns	essa	me	quais	dito	ou	sei
cujo	estávamos	pequena	das	também	sob	amplo	vendo	essas	mesma	qual	diz	outra	onde
cujos	este	pequenas	de	tampouco	sobre	amplos	ver	esse	mesmas	quando	dizem	outras	vou
da	estes	pequeno	dela	te	sua	ante	vez	esses	mesmo	quanto	do	outro	tenho
das	estou	pequenos	delas	tem	suas	antes	vindo	esta	mesmos	quantos	dos	outros	quero
de	eu	per	dele	tendo	tal	ao	vir	estas	meu	que	e	pelas	
dela	fazendo	perante	deles	tenha	tambem	muito	vos	este	meus	ser	nem	pelas	
dele	fazer	pode	depois	ter	teu	muitos	vós	estes	minha	vai	nenhum	pelo	
deles	feita	pôde	dessa	teu	teus	na	aqui	ha	minhas	vamos	nessa	pelos	
demais	feitas	podendo	dessas	teus	toda	não	mim	isso	muita	então	nessas	perante	
depois	feito	poder	desse	ti	todas	nas	ser	isto	muitas	nada	nesta	pois	

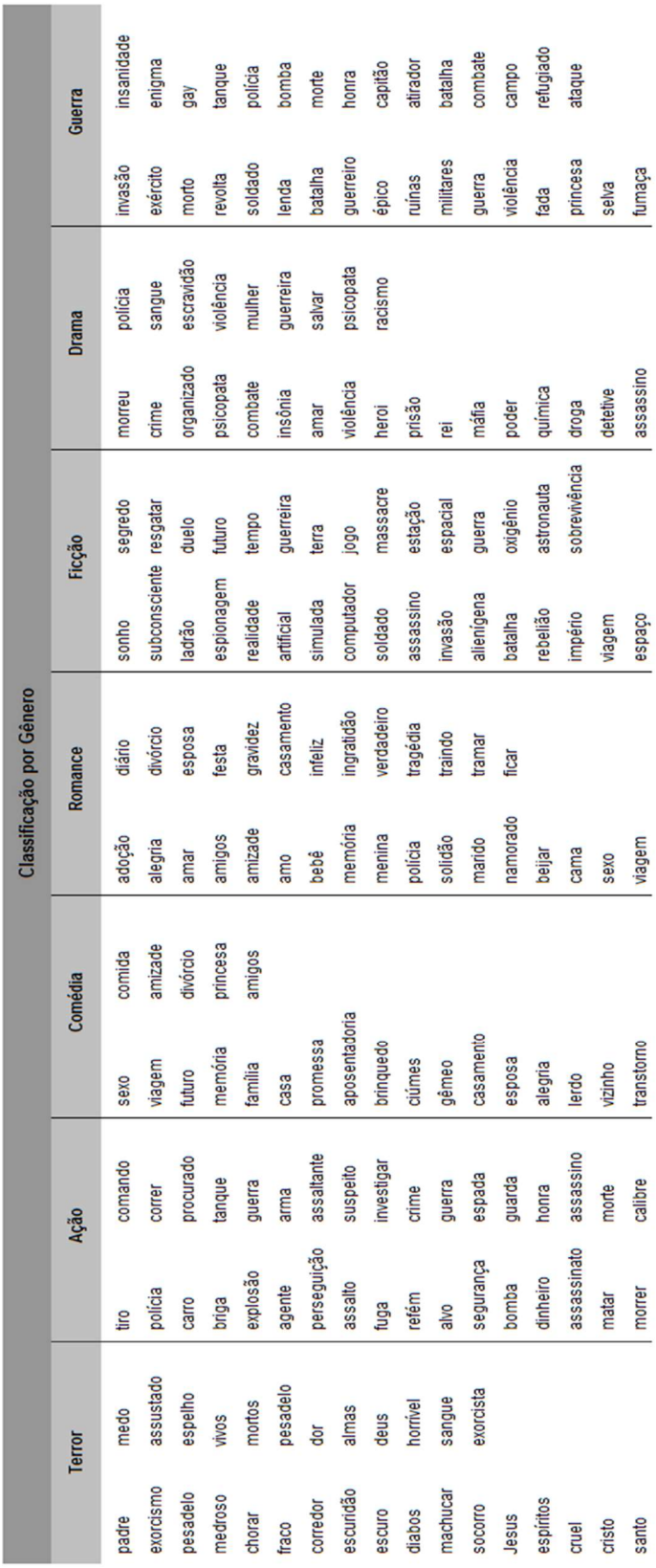
Figura 31: Lista de stop-words
Fonte: Elaborada pelo autor

APÊNDICE B – Classe Java

```
package lucene;
import org.apache.lucene.index.DirectoryReader;
public class indexar {
    static void indexDirectory() {
         try {
         Path path = Paths.get("/home/func/armstrong/Documentos/lucene/indice");
         Directory directory = FSDirectory.open(path);
         IndexWriterConfig config = new IndexWriterConfig(new BrazilianAnalyzer());
         IndexWriter indexWriter = new IndexWriter(directory, config);
         indexWriter.deleteAll();
         File f = new File("/home/func/armstrong/Documentos/lucene/arquivos"); // current directory
             for (File file : f.listFiles()) {
                    System.out.println("Indexando o arquivo - " + file.getCanonicalPath());
                    Document doc = new Document();
                    doc.add(new TextField("path", file.getName(), Store.YES));
                    FileInputStream is = new FileInputStream(file);
                    BufferedReader reader = new BufferedReader(new InputStreamReader(is));
                    StringBuffer stringBuffer = new StringBuffer();
                    String line = null;
                    while((line = reader.readLine())!=null){
                      stringBuffer.append(line).append("\n");
                    }
                    reader.close();
                    doc.add(new TextField("contents", stringBuffer.toString(), Store.YES));
                    indexWriter.addDocument(doc);
             }
             indexWriter.close();
             directory.close();
         } catch (Exception e) { e.printStackTrace(); }
    }
    static String[] search(String text) {
        String vetor[] = new String[50];
        try {
            Path path = Paths.get("/home/func/armstrong/Documentos/lucene/indice");
            Directory directory = FSDirectory.open(path);
            IndexReader indexReader =  DirectoryReader.open(directory);
            IndexSearcher indexSearcher = new IndexSearcher(indexReader);
            QueryParser queryParser = new QueryParser("contents",  new BrazilianAnalyzer());
            Query query = queryParser.parse(text);
            TopDocs topDocs = indexSearcher.search(query,10);
            System.out.println("totalHits " + topDocs.totalHits);
            int cont=0;
            for (ScoreDoc scoreDoc : topDocs.scoreDocs) {
               Document document = indexSearcher.doc(scoreDoc.doc);
               System.out.println("INFO BUSCADOR - Caminho:" + document.get("path") + " - Score: " + scoreDoc.score);
               vetor[cont]="INFO BUSCADOR - Caminho:" + document.get("path") + " - Score: " + scoreDoc.score;
                cont++;
            }
        } catch (Exception e) { e.printStackTrace(); }
        return (vetor);
    }
 }
```

Figura 32: Classe Java
Fonte: Elaborada pelo autor

APÊNDICE C – “Tabela base” de gêneros



Classificação por Gênero																					
Terror		Ação		Comédia		Romance		Ficção		Drama		Guerra	
padre	medo	tiro	comando	sexo	comida	adoção	diário	sonho	segredo	morreu	polícia	invasão	insanidade
exorcismo	assustado	polícia	correr	viagem	amizade	alegria	divórcio	subconsciente	resgatar	crime	sangue	exército	enigma
pesadelo	espelho	carro	procurado	futuro	divórcio	amar	esposa	ladrão	duelo	organizado	escravidão	morto	gay
medroso	vivos	briga	tanque	memória	princesa	amigos	festa	espionagem	futuro	psicopata	violência	revolta	tanque
chorar	mortos	explosão	guerra	família	amigos	amizade	gravidez	realidade	tempo	combate	mulher	soldado	polícia
fraco	pesadelo	agente	arma	casa		amo	casamento	artificial	guerreira	insônia	guerreira	lenda	bomba
corredor	dor	perseguição	assaltante	promessa		bebê	infeliz	simulada	terra	amar	salvar	batalha	morte
escuridão	almas	assalto	suspeito	aposentadoria		memória	ingratidão	computador	jogo	violência	psicopata	guerreiro	honra
escuro	deus	fuga	investigar	brinquedo		menina	verdadeiro	soldado	massacre	heroi	racismo	épico	capitão
diabos	horrível	refém	crime	ciúmes		polícia	tragédia	assassino	estação	prisão		ruínas	atirador
machucar	sangue	alvo	guerra	gêmeo		solidão	traindo	invasão	espacial	rei		militares	batalha
socorro	exorcista	segurança	espada	casamento		marido	tramar	alienígena	guerra	máfia		guerra	combate
Jesus		bomba	guarda	esposa		namorado	ficar	batalha	oxigênio	poder		violência	campo
espíritos		dinheiro	honra	alegria		beijar		rebelião	astronauta	química		fada	refugiado
cruel		assassinato	assassino	lerdo		cama		império	sobrevivência	droga		princesa	ataque
cristo		matar	morte	vizinho		sexo		viagem		detetive		selva	
santo		morrer	calibre	transtorno		viagem		espaço		assassino		fumaça	

Figura 33: Palavras-chave por gênero
Fonte: Elaborada pelo autor

APÊNDICE D – "Tabela base" de sentimentos

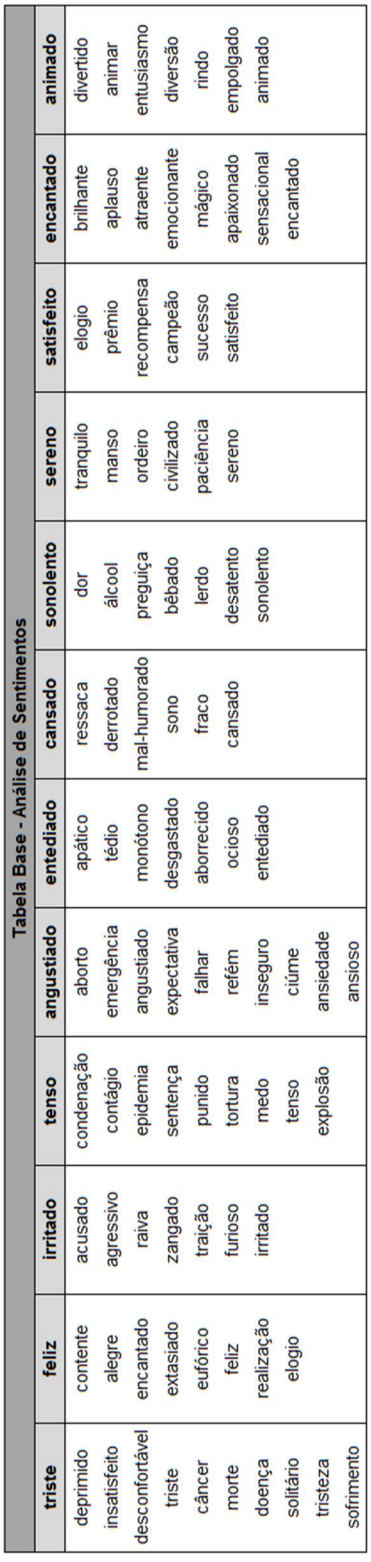


Tabela Base - Análise de Sentimentos											
triste	**feliz**	**irritado**	**tenso**	**angustiado**	**entediado**	**cansado**	**sonolento**	**sereno**	**satisfeito**	**encantado**	**animado**
deprimido	contente	acusado	condenação	aborto	apático	ressaca	dor	tranquilo	elogio	brilhante	divertido
insatisfeito	alegre	agressivo	contágio	emergência	tédio	derrotado	álcool	manso	prêmio	aplauso	animar
desconfortável	encantado	raiva	epidemia	angustiado	monótono	mal-humorado	preguiça	ordeiro	recompensa	atraente	entusiasmo
triste	extasiado	zangado	sentença	expectativa	desgastado	sono	bêbado	civilizado	campeão	emocionante	diversão
câncer	eufórico	traição	punido	falhar	aborrecido	fraco	lerdo	paciência	sucesso	mágico	rindo
morte	feliz	furioso	tortura	refém	ocioso	cansado	desatento	sereno	satisfeito	apaixonado	empolgado
doença	realização	irritado	medo	inseguro	entediado		sonolento			sensacional	animado
solitário	elogio		tenso	ciúme						encantado	
tristeza			explosão	ansiedade							
sofrimento				ansioso							

Figura 34: Palavras-chave por sentimentos baseada na tese de Osiek (2014) e no artigo de Mohammad e Turney (2010)
Fonte: Elaborada pelo autor