Proposta de Modelo de Recomendação de Conteúdo baseado em Arquivos de Legendas de Filmes e Séries

Armstrong Gomes Brito
Universidade FUMEC, Email: armstrongg1@gmail.com

Luiz Claudio Gomes Maia
Universidade FUMEC, Email: luizclaudiomaia@gmail.com

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 Lucene* 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 apresentado 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 relevantes, 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.

1 INTRODUÇÃO

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. 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 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. 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. 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. As ferramentas de recuperação da informação bem como os sistemas de recomendação devem ajudar o usuário, ou seja, devem ser intuitivas e não os confundi-los.

Diante deste cenário, esta pesquisa[1] demonstra 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). A proposta deste trabalho é justamente ampliar a possibilidade de busca e garantir que o usuário tenha uma opção de ferramenta para encontrar justamente títulos que sejam relevantes.

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 da obra, é ele que será apresentado ao usuário para consequentemente 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. Os metadados têm sua parcela de importância, mas os sistemas de recuperação da informação devem enfatizar o conteúdo.

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. A utilização dos arquivos de legendas mencionados são os arquivos com extensão .srt (produzidos em diferentes línguas), podem atender a esta expectativa, porque é um arquivo pequeno, de poucos *bytes*, que pode ser processado e recuperado rapidamente mesmo que o usuário esteja distante do servidor.

O mercado pode direcionar suas estratégias muito além das buscas por padrões. O usuário deve ter novas opções de escolhas e sistemas de recomendação e filtro de conteúdo eficazes. Novos filmes e séries podem ser criadas de acordo com que os usuários procuram e os sistemas de recomendação podem ajudar neste caminho. O usuário não pode ser induzido a

[1] Apoio FAPEMIG e PROPIC/FUMEC

preencher inúmeros formulários, ter um cadastro e perfil completo, além da necessidade de realizar avaliações para que encontre o que esteja procurando.

A contribuição deste trabalho será propor um modelo contemplando: pesquisa, classificação por gênero, análise de sentimentos e a identificação de títulos similares. A busca por padrões e pelos metadados são realizadas por inúmeros software, mas algumas pesquisas simples e práticas, como as indicadas neste trabalho, por conteúdo, devem ser uma alternativa para complementar as lacunas destes sistemas.

2 RECUPERAÇÃO DA INFORMAÇÃO

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.

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.

Um dos diagramas que descrevem o processo de recuperação de informação em sistemas é o de Cardoso (2000) apresentado na Figura 1, 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 1 – Exemplo de sistema de recuperação da informação

Fonte: Gey (1992).

Dentro do tema da recuperação da informação temos os sistemas de recuperação da informação. Para Souza *et al*. (2006 apud LANCASTER, 1968), os sistemas de recuperação da informação-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 *software*.

A maior parte dos sistemas de recuperação de informação desenvolvidos hoje em dia são concebidos 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.

A análise de sentimentos é um 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, p.43)

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. 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). Neste artigo foi desenvolvido um artefato (seção 4) para protagonizar a implementação prática do modelo.

Primeiro é apresentado as ferramentas *Apache Lucene* e *OGMA* que serão utilizadas na proposta do modelo. Em seguida, é abordada a especificação do modelo, que será dividido em quatro etapas: passos iniciais (requisitos); pesquisa; classificação por gênero e sentimentos; e identificação de títulos similares. Por último, um resumo do modelo é apresentado.

3.1 FERRAMENTAS/SOFTWARE PARA CONSTRUÇÃO DO MODELO

O *Apache Lucene* é uma ferramenta livre utilizada para a recuperação da informação em arquivos textuais. *Lucene* é uma biblioteca de mecanismo de procura de texto altamente escalável e de software livre a partir do Apache Software Foundation. Para este trabalho foi utilizado a versão 5.2.1 em conjunto com o *Eclipse*. O *Eclipse* é uma IDE (*Integrated*

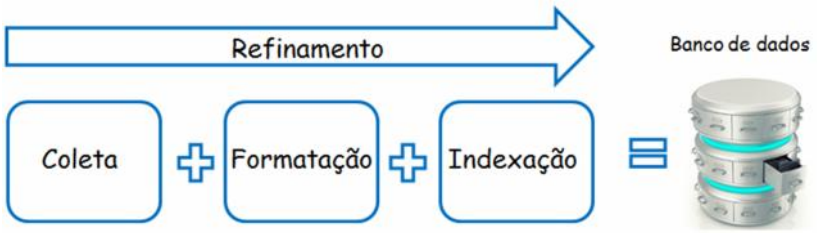
Development Environment) para desenvolvimento *Java*, porém suporta várias outras linguagens. Ele segue o modelo open source de desenvolvimento de *software*. A versão utilizada do *Eclipse* é a *Mars Release* (4.5.0). Neste trabalho o *Apache Lucene* será utilizado para indexar os arquivos de legenda formatados para em seguida possibilitar a implementação da busca/pesquisa dos arquivos indexados.

Outro *software* utilizado foi o *OGMA* (http://www.luizmaia.com.br/ogma/) que 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. 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*".

3.2 PASSOS INICIAIS: COLETA, FORMATAÇÃO E INDEXAÇÃO

Estes passos iniciais 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 2 resume estes passos finalizando a criação do banco de dados. Cada resultado gerado em uma etapa é utilizado na próxima.

Figura 2 - Requisitos do modelo - Primeira etapa



Fonte: Elaborada pelos autores.

O primeiro passo consiste em levantar e adquirir todos os arquivos de legendas que serão utilizados. Os arquivos de legendas mencionados são os arquivos com extensão .srt que são produzidos em diferentes línguas.

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. Foi necessário preparar todos os arquivos retirando e refinando os dados que serão indexados. Finalmente, cada arquivo, deverá conter estritamente as falas dos personagens.

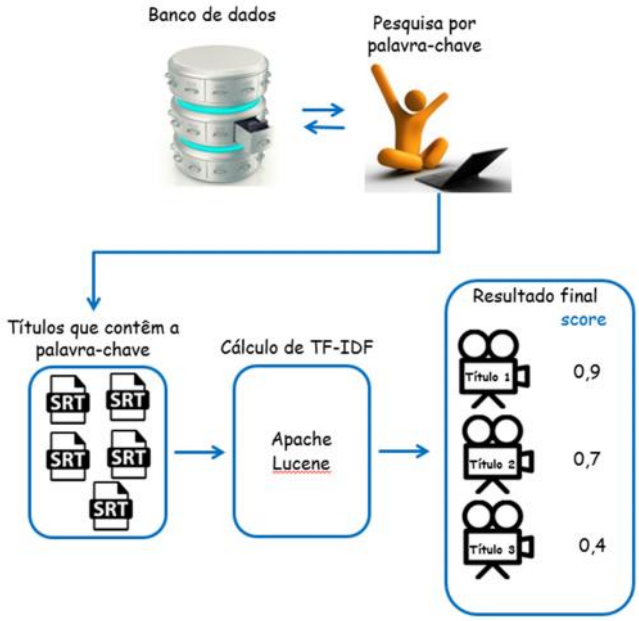
O último passo para construção do banco de dados visa possibilitar a construção da proposta do modelo de recuperação da informação em arquivos de legendas é a indexação. A indexação é necessária para permitir a rápida recuperação dos dados indexados, além de possibilitar a análise dos termos que mais representam cada filme ou série. A indexação será realizada através do *Apache Lucene*.

3.3 PESQUISA POR PALAVRAS-CHAVE

Após o desenvolvimento dos requisitos, iniciou-se a segunda etapa do modelo. 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 consistiu na especificação da uma interface de pesquisa de filmes e séries já indexados na etapa anterior.

O procedimento de pesquisa por palavra-chave é resumido pela figura 3. O usuário insere o termo que deseja pesquisar no banco de dados. O resultado será uma lista ordenada dos documentos mais relevantes, ou seja, apresentará um ranking dos documentos que melhor representam aquele termo. A busca e o ranqueamento são realizados pelo *Apache Lucene*.

Figura 3 – Proposta de modelo - Segunda etapa



Fonte: Elaborada pelos autores.

3.4 CLASSIFICAÇÃO POR GÊNERO

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.

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”, composta por termos que identificarão cada gênero, é o primeiro passo desta etapa.

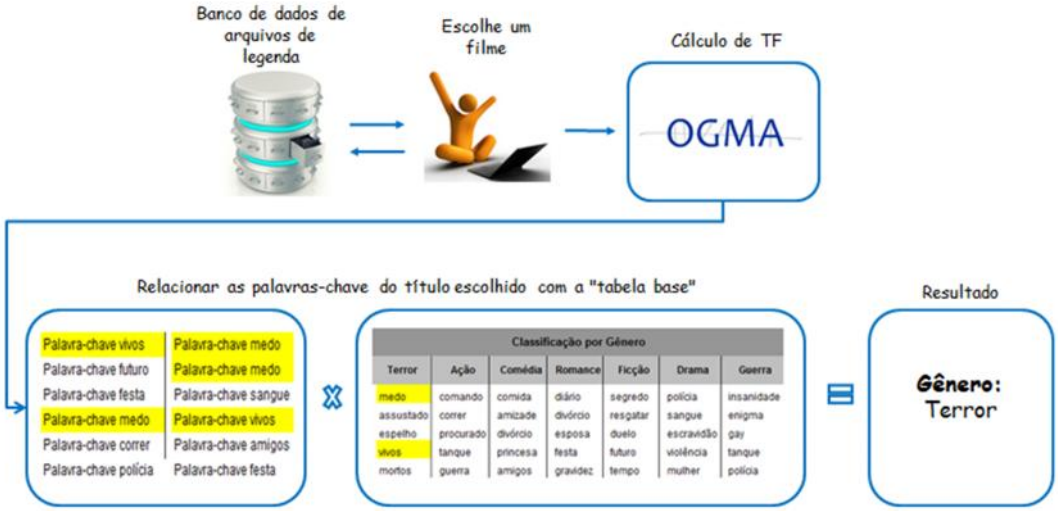
A “tabela base” foi construída da seguinte forma:

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. Por meio 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;

Com a “tabela base” criada pode-se iniciar o processo de classificação por gênero (resumido pela figura 4). Primeiro o usuário escolhe o título que deseje categorizar e utiliza-se o *OGMA* (opção extrair termos sem *stop-words*) para a identificação das dez palavras-chave do mesmo. As dez primeiras palavras-chave é que serão utilizadas. Na próxima etapa é feita a relação das palavras-chave com os termos da “tabela base” criada para a classificação. O número de palavras-chave que mais estiverem relacionadas com uma única categoria, classificará o título. A figura 4, apresenta como exemplo, as palavras-chave “vivos” e “medo”, logo a classificação de gênero para o respectivo filme será Terror.

Figura 4 – Proposta de modelo - Terceira etapa



Fonte: Elaborada pelos autores.

3.4.1 Análise de sentimentos

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

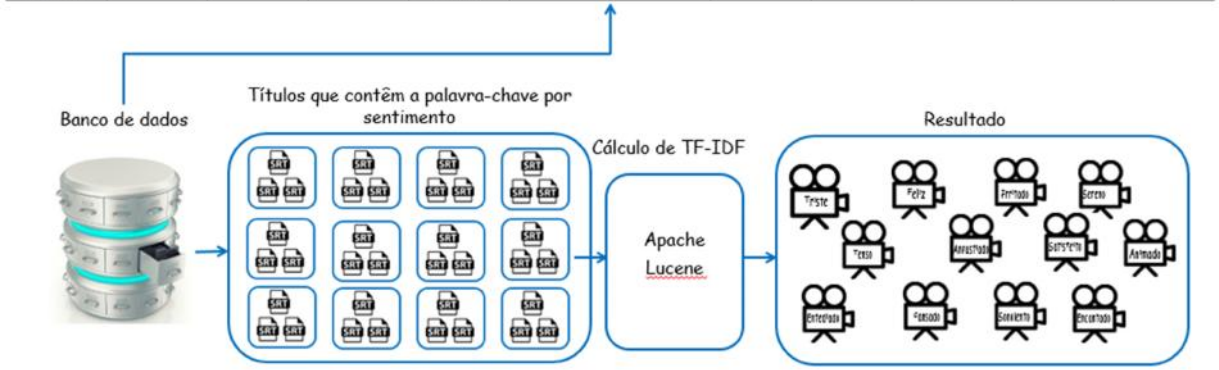
• triste
• feliz
• irritado
• tenso
• angustiado
• entediado
• cansado
• sonolento
• sereno
• satisfeito
• encantado
•animado

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 foi baseada em nos trabalhos de Osiek (2014), na qual ele propõe um modelo linguístico emocional apresentando as palavras-chave para cada sentimento e no de Mohammad e Turney (2010), que apresenta um método léxico, denominado NRC *Emoticon Lexicon*, que classifica textos em oito categorias afetivas.

A figura 5 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 de legendas, 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 (dada pelo Lucene) de sua pontuação final, será o que melhor representará cada sentimento. Basicamente, será realizado a tentativa de captar qual filme ou série transmite melhor um sentimento específico.

Figura 5 – Proposta de adaptação do modelo - análise 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							



Fonte: Elaborada pelos autores.

3.5 IDENTIFICAÇÃO DE TÍTULOS SIMILARES

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 encontrada em sistemas de recomendação do tipo.

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;

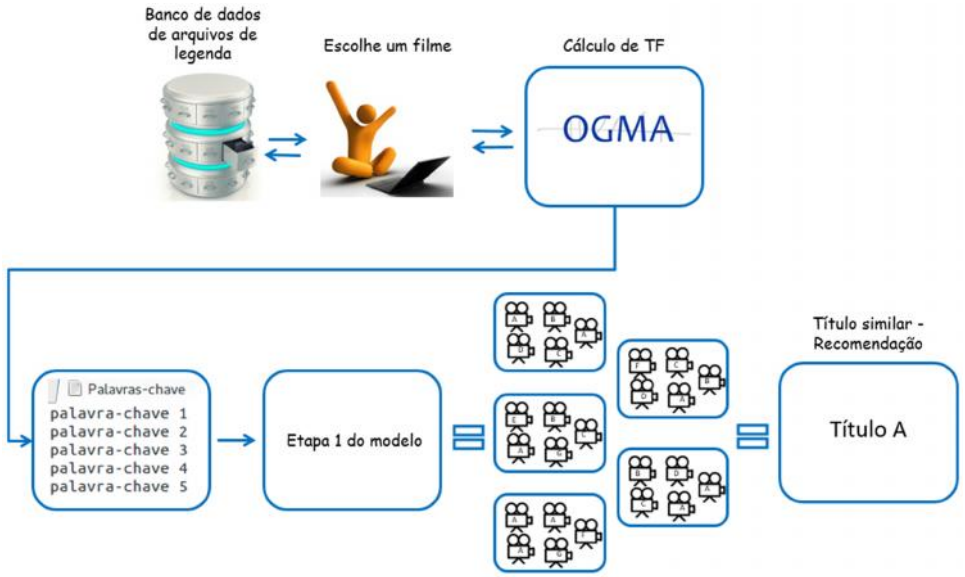
• Estas cinco palavras-chave serão utilizadas como parâmetro de pesquisa para identificar os cinco primeiros títulos do ranqueamento (via Apache *Lucene*) para cada uma;

• Os títulos recomendados serão baseados nos que aparecem com mais frequência, respeitando o título que apresentar a maior soma da pontuação final;

A figura 6 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, serão utilizadas 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 6 – Proposta de modelo - Quarta etapa

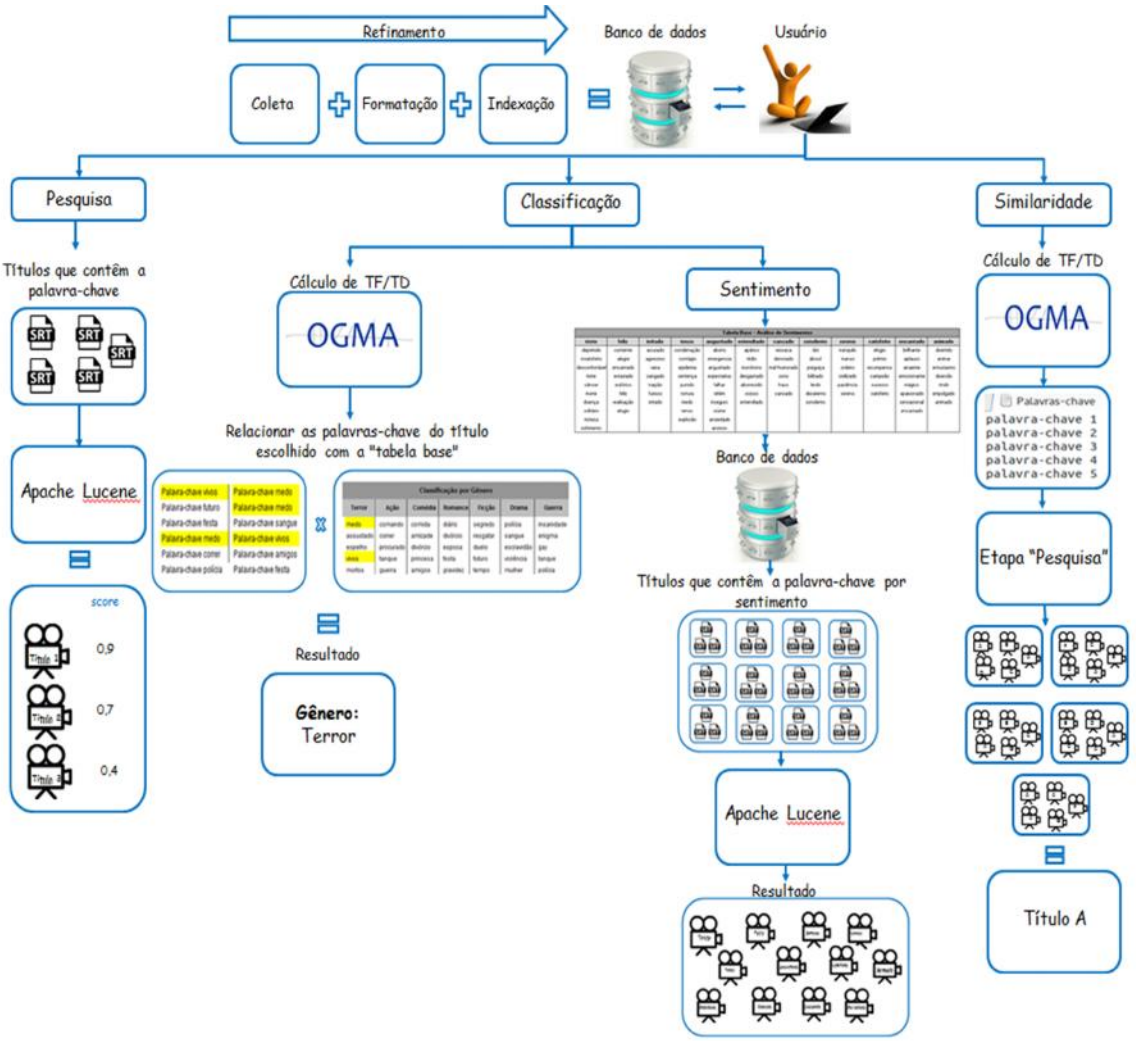


Fonte: Elaborada pelos autores.

3.6 RESUMO DO MODELO

Todo o procedimento para idealização do modelo é resumido na figura 7. 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. O modelo foi dividido em quatro etapas para melhor visualização e entendimento. Ainda na figura 7, é demonstrado as etapas de Pesquisa, Classificação, Análise de sentimentos e Similaridade. A proposta deste trabalho foi apresentar como resultado um modelo de busca e recomendação de conteúdo.

Figura 7 – Overview do modelo



Fonte: Elaborada pelos autores.

4 ARTEFATO: IMPLEMENTAÇÃO E TESTE DO MODELO

Para testar o modelo proposto foram demonstrados alguns exemplos práticos e também algumas análises comparativas com o site IMDb. Todas as validações aqui apresentadas foram idealizadas e executadas pelos autores deste trabalho. 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.

4.2 BUSCA POR PALAVRAS-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. Optou-se em efetuar algumas buscas utilizando três palavras-chave que o site IMDb identifica para cada título. 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 e questão. 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 figura 8 expressa bem os resultados encontrados. O nível de aceitação da categorização dos títulos por legendas foi alto em comparação com a classificação realizada pelo site IMDb. 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.

Figura 8 – Compilação dos resultados - segunda etapa

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%

Fonte: Elaborada pelos autores.

4.3.1 Análise de sentimentos

O resultado para cada sentimento é demonstrado na figura 9. 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.

Figura 9 – Compilação dos resultados - análise de sentimentos

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

Fonte: Elaborada pelos autores.

4.4 TÍTULOS SEMELHANTES

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;

5 CONSIDERAÇÕES FINAIS

Hoje o acervo de dados e informações são imensos, 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 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.

Com o modelo, verificou-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 conteúdo 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.

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 possibilita a pesquisa em cima de conteúdo e não apenas de dados parametrizados. Este trabalho contribui para a área da Ciência da Informação, que tem trabalhado a questão da recomendação de conteúdo e das possibilidades geradas pelo uso destas ferramentas. Uma destas possibilidade foi demonstrada neste trabalho com 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, que se identifica com a ciência da informação, pois ela estuda a informação desde o seu início até o processo de transformação de dados em conhecimento.

O estudo realizado apresentou algumas limitações: primeiro quanto a coleta de dados;

segundo 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.

Para trabalhos futuros, 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.

Finalizando, é indicado 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.

Motion for Content Recommendation Model-Based File Movies and Legends Series

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 relevant recommendations, as they provide the user freedom to search within a specific content. Already the gender sorting 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 sorting. Apache-Lucene. OGMA. Recommender systems.

REFERÊNCIAS

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.

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>.

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

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

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

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.

GEY, F. F. Models in Information Retrieval. Folders of Tutorial Presented at the 19th ACM **Conference on Research and Development in Information Retrieval**. [S.l.], 1992. Folder.

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 J**ournal on Computing (JoC), Global Science and Technology Forum**, v. 4, n. 3, p. 13, 2015.

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

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.

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

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

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

OSIEK, B. A. **Reconhecimento de sentimento em texto abordado através da computação afetiva**. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2014.

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.

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

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.

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

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.