THIAGO BELLOTTI FURTADO

ABORDAGEM HÍBRIDA DE RECOMENDAÇÃO DE CONTEÚDO BASEADA EM TAGS ADAPTATIVAS APLICADA EM BIBLIOTECAS DIGITAIS

LAVRAS – MG

2016

THIAGO BELLOTTI FURTADO

ABORDAGEM HÍBRIDA DE RECOMENDAÇÃO DE CONTEÚDO BASEADA EM TAGS ADAPTATIVAS APLICADA EM BIBLIOTECAS DIGITAIS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Banco de Dados e Engenharia de Software, para a obtenção do título de Mestre.

Prof. Dr. Ahmed Ali Abdalla Esmin

Orientador

LAVRAS – MG

2016

Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos

da Biblioteca Universitária da UFLA

Furtado, Thiago Bellotti

ABORDAGEM HÍBRIDA DE RECOMENDAÇÃO DE CONTEÚDO BASEADA EM TAGS ADAPTATIVAS APLICADA EM BIBLIOTECAS DIGITAIS / Thiago Bellotti Furtado. 1^a ed. rev., atual. e ampl. – Lavras : UFLA, 2016.

76 p. : il.

Dissertação(mestrado)–Universidade Federal de Lavras, 2016.

Orientador: Prof. Dr. Ahmed Ali Abdalla Esmin.

Bibliografia.

1. Sistemas de Recomendação. 2. Repositório institucional. 3. Biblioteca Digital. 4. Tags. 5. Recuperação da Informação.

CDD-808.066

THIAGO BELLOTTI FURTADO

ABORDAGEM HÍBRIDA DE RECOMENDAÇÃO DE CONTEÚDO BASEADA EM TAGS ADAPTATIVAS APLICADA EM BIBLIOTECAS DIGITAIS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Banco de Dados e Engenharia de Software, para a obtenção do título de Mestre.

APROVADA em 22 de Agosto de 2016.

Prof. Dr. Ahmed Ali Abdala Esmin	UFLA
Prof. Dr. Carlos Henrique Valério de Moraes	UFNIFEI
Prof. Dr. Wilian Soares Lacerda	UFLA

Prof. Dr. Ahmed Ali Abdalla Esmin
Orientador

LAVRAS – MG
2016

Dedico esse trabalho aos meu pais, Gonzaga e Luciana,
e minhas irmãs, Tatiana e Thaís.

AGRADECIMENTOS

Aos meus pais e irmãs, pelo apoio incondicional, que mesmo distantes sempre estiveram presentes, com palavras de incentivo e motivação nos momentos mais difíceis.

Agradeço ao meu orientador Prof. Ahmed pelas orientações e conselhos que agregaram novas perspectivas de trabalho e conhecimento. Também pela amizade e confiança concedida para realizar essa pesquisa. A todos os professores do programa de mestrado pelos ensinamentos, em especial, Prof. André Zambalde, Prof. Luiz Henrique e Prof. Wilian Lacerda pelas sugestões e críticas construtivas que agregaram nesse trabalho.

A todos os colegas da Biblioteca Universitária da UFLA em especial ao Vicente Terra, pelo auxílio e empenho dedicado na construção desse projeto.

A todos os colegas discentes do mestrado pela oportunidade de convivência e experiências compartilhadas, em especial, Bruno Rezende, Ednaldo Oliveira, Fernando Simeone e João Antonio, pela colaboração durante o período acadêmico.

A todos os amigos de Guarani e Lavras pelos momentos de descontração que tornaram os dias menos apreensivos.

Aos meus avós e a Geny Barbosa e toda família, pelos conselhos e orações. A Deus, por proporcionar-me tal oportunidade, de seguir com saúde e determinação para desenvolver meu projeto de mestrado.

A todos que de alguma forma colaboraram para a realização desse trabalho. Obrigado!

Tudo parece impossível até que seja feito.
(Nelson Mandela)

RESUMO

Com a evolução tecnológica das bibliotecas no ambiente acadêmico, grande quantidade de informações e documentos são disponibilizados para acesso, mas nem sempre esses sistemas possuem mecanismos capazes de buscar de forma integrada informações relevantes para o usuário. Para amenizar este problema, propomos um sistema de recomendação que gera o perfil do usuário por meio de tags que são remodeladas ao longo do tempo. Para traçar o perfil do usuário o sistema utiliza informações do seu histórico de empréstimos armazenado na base de dados da biblioteca e coleta suas opiniões (feedback) por meio de uma lista de recomendações. Esses dados são integrados com a base de documentos do repositório institucional. Desta forma, o sistema de recomendação auxilia os usuários na identificação de itens relevantes e faz sugestões de conteúdo em um ambiente integrado que contem documentos do repositório institucional e da base de dados da biblioteca da universidade. O sistema de recomendação proposto utiliza uma abordagem híbrida sendo aplicado em um ambiente acadêmico com a participação dos usuários.

Palavras-chave: Sistemas de Recomendação, Repositório institucional, Biblioteca Digital, Tags, Recuperação da Informação.

ABSTRACT

The technological evolution of the library in the academic environment brought a lot of information and documents that are available to access, but not always these systems have mechanisms to search in an integrated way the relevant information for the user. To alleviate this problem, we propose a recommendation system that generates the user profile through tags that are reshaped over time. To trace the user profile the system uses information from your lending history stored in the library database and it collects their opinions (feedback) through a list of recommendations. These data are integrated with the document base of institutional repository. Thus, the recommendation system assists users in identifying relevant items and makes suggestions for content in an integrated environment that contains institutional repository documents and the university library database. The proposed recommendation system uses a hybrid approach being applied in an academic environment with the participation of users.

Keywords: Recommender Systems, Institutional Repository, Digital Library, Tags, Information Retrieval.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

As modificações tecnológicas e as recentes concepções de gerenciamento de recursos de informação têm causado uma quebra no paradigma dos modelos tradicionais de bibliotecas. O conceito de bibliotecas digitais apresenta uma alternativa para ampliar as condições de busca, disponibilidade e recuperação de informações de maneira globalizada, principalmente baseando-se sobre redes de acesso a web.

O avanço nesse tipo de tecnologia proporcionou às pessoas e organizações, acesso a uma grande quantidade de documentos e informações. Porém, na maioria das vezes esse potencial não é bem explorado. Isso ocorre devido há alguns fatores complicadores, como a falta de integração entre os sistemas que não possuem mecanismos para facilitar a busca de novas informações (TEJEDA-LORENTE et al., 2014).

Por armazenar grande volume de documentos e informações que são disponibilizados a toda comunidade acadêmica, os sistemas de bibliotecas digitais das universidades, também conhecidas como Repositórios Institucionais (RI) tornaram-se uma importante ferramenta para a democratização da produção científica, apoiando a aprendizagem, ensino e pesquisa. Portanto, tem-se adotado repositórios institucionais como meio para disseminar, armazenar e preservar as informações a fim de permitir o acesso permanente e confiável da produção científica. Tem sido amplamente adotada nas universidades em âmbito nacional e internacional para auxiliar na promoção do acesso à informação de forma transparente e democrática, aumentando o impacto do reconhecimento e da visibilidade nacional e internacional das produções científicas desenvolvidas nas instituições (MEDEIROS, 2012).

Nesse contexto, são identificados sistemas que atendem a demandas similares e possuem estruturas de armazenamento com dados semelhantes. Como nos repositórios institucionais, os sistemas de gestão de bibliotecas armazenam em sua base de dados os metadados que contem informações como, título do livro, nome do autor do livro, palavras chave que classificam o conteúdo de um livro, entre outros. Esses sistemas armazenam um histórico de todas as atividades realizadas por seus usuários e podem ser úteis para tentar identificar o perfil e o interesse de cada usuário inserido no ambiente acadêmico. No entanto, torna-se essencial que exista uma integração capaz de facilitar a busca por conteúdo de forma padronizada e simultânea em ambos os sistemas.

A expansão das bases de dados de livros e documentos digitais nas bibliotecas contribuiu para o aumento da diversidade de temas, acarretando em problemas como o de sobrecarga de informações. Geralmente o usuário não tem percepção de todo conteúdo que é disponibilizado para acesso, ou então simplesmente busca somente conteúdo que seja de seu conhecimento, mas não exploram outros assuntos disponíveis que possam ser de seu interesse.

Com objetivo de auxiliar usuários na busca por informações pertinentes, pesquisas abordaram temas semelhantes, como é o caso de sistemas de marcação social, em que usuários expressam suas preferências compartilhando marcações (tagging). As tags podem ser consideradas palavras curtas que transmitem alguma informação e são utilizadas para identificar interesse dos usuários, servindo como mecanismo para gerar recomendações personalizadas (DURAO; DOLOG, 2012).

Um sistema de recomendação pode combinar essas informações para auxiliar os usuários na identificação eficaz de itens adequados a sua necessidade ou preferência, orientando-os de forma personalizada ao sugerir itens relevantes entre uma grande quantidade de opções. Assim, evita-se a sobrecarga de informações irrelevantes ao filtrar o acesso à informação de usuários que não tem conhecimento detalhado sobre o que desejam encontrar. Geralmente, esses sistemas coletam in-

formações históricas produzidas ao longo do tempo por certos usuários, para tentar traçar seu perfil com base nas suas preferências (BARRAGáNS-MARTÍNEZ et al., 2010).

As técnicas aplicadas em sistemas de recomendação são classificadas basicamente em duas formas: filtragem colaborativa e filtragem baseada em conteúdo. A filtragem colaborativa (FC) considera um grupo de usuários para fazer recomendações conforme os interesses existentes de acordo com certa semelhança dentro do grupo. Para que essa técnica seja aplicada, um usuário deve fazer avaliações de determinado item. Caso exista alguma semelhança entre os usuários, um item que foi avaliado poderá ser recomendado sem a necessidade de avaliação por parte do usuário que irá receber a recomendação (GARCIA; FROZZA, 2013). Em contrapartida, a técnica de filtragem baseada em conteúdo depende que o usuário avalie um item para receber recomendações semelhantes ao item avaliado, dessa forma não há dependência das preferências de outros usuários similares e as recomendações são direcionadas especificamente para um usuário de acordo com a similaridade dos itens de seu interesse (LIU et al., 2015).

As duas técnicas de recomendação descritas podem auxiliar usuários na escolha de determinado conteúdo, porém ainda apresentam pontos fracos que podem ser desfavoráveis para o processo de recomendação. A principal desvantagem do sistema de filtragem baseado em conteúdo está relacionada com o excesso de especialização por parte de um usuário em relação a um item, isso significa que um usuário tende a escolher sempre itens semelhantes ao que tinha optado antes. Itens que não possuem as características dentro dos padrões de escolha de um usuário poderão ser descartados, mas se fossem recomendados como itens inesperados, poderiam despertar o interesse do usuário (KIM et al., 2010).

A abordagem FC também apresenta algumas limitações: problema relacionado a dados esparsos, arranque a frio e escalabilidade, abordados em Durao e Dolog (2012), Kim et al. (2010), Ji e Shen (2015) e Polatidis e Georgiadis (2016).

O problema de dados esparsos ocorre quando os dados disponíveis são insuficientes para identificar usuários ou itens semelhantes, devido a grande quantidade de usuários e itens. Isso pode ocorrer quando muitos usuários classificam poucos itens, ou então itens muito populares foram classificados por poucos usuários. Mesmo sendo possível o cálculo da semelhança, este não se torna confiável devido a insuficiências de informação para serem processadas (POLATIDIS; GEORGIADIS, 2016). O arranque a frio é um problema causado pela entrada de novos usuários no sistema que ainda não realizaram qualquer tipo de avaliação, o que impossibilita a geração de recomendações até que algum item seja classificado.

Em (POLATIDIS; GEORGIADIS, 2016) divide esse problema em itens de arranque a frio, quando se trata de um novo item que ainda não foi avaliado, e usuários de arranque a frio, quando são inseridos novos usuários que não fizeram avaliações de itens. A falta de escalabilidade é uma limitação que pode afetar a qualidade das recomendações, nesse caso o número de usuários e itens aumenta ao longo do tempo, impossibilitando recomendações para novos itens e usuários, até que haja classificações e avaliações entre eles (DURAO; DOLOG, 2012).

Esses problemas apresentados podem ser amenizados combinando as técnicas de FC com a baseada em conteúdo, resultando em uma abordagem conhecida como híbrida. Ao utilizar essa abordagem, as duas técnicas são combinadas para melhorar a precisão das recomendações, reduzindo as desvantagens e aumentando os benefícios (LIU et al., 2015). Dessa forma, quando não existem informações dos usuários e suas avaliações, aplicam-se mecanismos usados em sistemas com filtragem baseado em conteúdo. Caso não existam informações suficientes sobre o conteúdo associado aos itens, então são aplicadas as técnicas usadas nos sistemas de filtragem colaborativa (KARDAN; EBRAHIMI, 2013).

Ainda assim, a utilização dessas técnicas em ambientes em que usuários alteram suas preferências constantemente não é suficiente para garantir um adequado sistema de recomendação. Em um ambiente universitário, a comunidade

acadêmica possui perfis mais dinâmicos, principalmente por parte dos alunos, que estão sujeitos a pesquisarem uma diversidade de temas periodicamente. Por isso, é preciso uma abordagem capaz de adaptar as recomendações de acordo com novas circunstâncias. Para tentar amenizar esse tipo de problema, aplica-se o conceito de detecção de novidade conforme Gama (2010), que torna possível reconhecer um conceito como novo e indicar o surgimento de novos conceitos de acordo com as mudanças ocorridas ao longo do tempo.

1.2 Problema e Proposta de Trabalho

A crescente adoção de sistemas de bibliotecas digitais em instituições de ensino e pesquisa vem aumentando gradativamente e tem proporcionado aos usuários acesso a uma grande quantidade de conteúdo diversificado. Apesar de serem ferramentas importantes para a disponibilização e a gestão de conteúdos digitais, ainda não possuem em sua estrutura nativa uma forma de busca inteligente capaz de associar os interesses do usuário com o conteúdo a ser recuperado.

Estudos como o de (TEJEDA-LORENTE et al., 2014) utiliza lógica fuzzy em um ambiente de bibliotecas digitais para melhorar a qualidade das recomendações de itens, enquanto que o estudo feito por (BARRAGáNS-MARTÍNEZ et al., 2010) utiliza uma metodologia hibrida com base em tags para recomendar filmes. No entanto, não há sistema que recomende conteúdo aos usuários de forma integrada, entre diferentes bases de dados digitais e a partir da extração de tags que possam representar o perfil dos usuários sugerindo conteúdo novo e relevante ao longo do tempo.

Por isso, nesse trabalho propõe-se um mecanismo para aprimorar e facilitar a exploração do conteúdo pelos usuários sobre grande quantidade de informações. Com base no empréstimo de livros feito por usuários, metadados são extraídos e transformados em tags que formam listas para identificar as preferências de cada usuário. Por meio dessas listas de tags as recomendações são geradas e en-

viadas para que o usuário possa avaliar, e enviar seu feedback que é utilizado para reformular as tags de cada lista. Isso proporciona novas sugestões de conteúdo que sejam mais precisas e estejam de acordo com o interesse do usuário. Aplica-se o conceito de detecção de novidade para adaptar as recomendações ao perfil do usuário de acordo com as alterações de interesse em determinado momento.

Para demonstrar a viabilidade da proposta, a abordagem desenvolvida é avaliada com medidas de precisão, recuperação e f-score. As sugestões de conteúdo fomentam a produção científica disponíveis nas bibliotecas digitais e auxiliam o usuário no processo de pesquisa sobre novo conteúdo, direcionando sua busca de modo a proporcionar maior visibilidade sobre o conteúdo disponível nos repositórios digitais, mas que podem não ter sido explorados.

1.3 Objetivos Gerais e Específicos

Este trabalho teve como objetivo o desenvolvimento de um mecanismo de recomendações de conteúdo baseado em listas de tags para minimizar o esforço de busca e facilitar a identificação de documentos relevantes aos usuários. Informações de empréstimos de livros e documentos digitais contidas nas bases de dados do sistema de gerenciamento da biblioteca e do repositório institucional foram integradas para formar o conjunto de dados da pesquisa. Os dados foram pré-processados e as técnicas de recomendações abordadas na literatura foram aplicadas para gerar listas de tags que representam o perfil do usuário a fim de permitir sugestões de conteúdo relevante. A exatidão das recomendações foram avaliadas por meio do feedback de usuários e a partir de métricas de validação.

Para cumprir esses objetivos, os seguintes objetivos específicos foram analisados:

1. Integração das bases de dados do repositório institucional e do sistema de gerenciamento de bibliotecas;

2. Coleta e tratamento das informações de metadados de histórico de empréstimos e de documentos digitais oriundas das bases;

3. Estudo e compreensão das técnicas de recomendação de conteúdo utilizando abordagens híbrida;

4. Estudo e compreensão de métricas utilizadas para validar a precisão e adaptabilidade as recomendações;

5. Desenvolvimento de um mecanismo para recomendar conteúdo baseado nas listas de tags extraídas dos metadados dos documentos digitais;

6. Avaliação das recomendações a partir de métricas e feedback dos usuários.

1.4 Tipos de Pesquisa

Uma pesquisa pode ser classificada quanto à sua abrodagem, natureza, objetivos e quanto aos seus procedimentos (JUNG, 2004).

Em relação à abordagem do problema, a pesquisa pode ser classificada como qualitativa ou quantitativa. Quando se trata de uma abordagem qualitativa, o foco não é a avaliação numérica do estudo, mas sim a profundidade da compreensão sobre um fato. Na abordagem quantitativa, é necessário recursos e técnicas estatísticas que traduzam em números os conhecimentos gerados pela pesquisa, sendo assim centrada na objetividade.

Quanto à natureza, a pesquisa pode ser classificada como básica, caso produza conhecimentos úteis para o desenvolvimento da ciência, porém sem aplicação prática prevista. Por outro lado, a pesquisa pode ser classificada como aplicada se os conhecimentos produzidos são aplicados na solução de problemas específicos.

Quando se trata dos objetivos a pesquisa pode ser exploratória caso promova a familiaridade com o problema, descritiva se as características de uma população ou fenômeno são expostas por meio da coleta de dados padronizada ou explicativa, se identifica fatores que causam ou influenciam um fenômeno.

A pesquisa pode ser classificada quanto aos procedimentos utilizados em sua execução. Sendo assim, pode ser uma pesquisa bibliográfica, quando concebida a partir de outros trabalhos publicados. Pode ser uma pesquisa documental, caso os materiais utilizados não recebam tratamento analítico. A pesquisa pode ser caracterizada como experimental se as variáveis definidas influenciam nos resultados. Uma pesquisa do tipo *survey* propõe o levantamento de conhecimento com interrogação direta de pessoas. Quando o pesquisador tem pouco controle sobre os fenômenos ou quando o objetivo é observar fatos inseridos em algum contexto real, classifica-se essa pesquisa como um estudo de caso. Na pesquisa *ex-post-facto*, os experimentos são realizados após a ocorrência dos fatos. A pesquisa classificada como pesquisa-ação, tem o objetivo de estabelecer uma relação com uma ação ou mesmo com um problema coletivo. Se os pesquisadores interagem com os cenários investigados a pesquisa pode ser classificada como pesquisa participante.

Sendo assim, essa pesquisa pode ser classificada em aplicada quanto à sua natureza, pois é fundamentada em *design science*, uma vez que utiliza conhecimentos estabelecidos para gerar novos conhecimentos com fins de aplicação, por meio do desenvolvimento de ferramentas necessárias para ações adequadas no domínio dos profissionais em seus campos de atuação (VAISHNAVI; KUECHLER, 2004). É uma pesquisa exploratória quanto a seu objetivo, pois promove a familiaridade com o problema. É uma pesquisa quantitativa quanto à abordagem do problema, pois os resultados obtidos são avaliados por meio de um conjunto de métricas que traduzem em números os conhecimentos gerados. Quanto aos procedimentos, é uma pesquisa do tipo *survey*, pois gera conhecimentos por meio da interrogação de pessoas,e também experimental, pois os parâmetros definidos influenciam nos resultados obtidos.

1.5 Estrutura do Trabalho

Este trabalho está dividido em 5 capítulos. No Capítulo 1 são abordados contextualização, motivação e proposta de trabalho, os objetivos gerais e específicos e o tipo de pesquisa utilizado nesse trabalho. O Capítulo 2 discorre sobre os assuntos e as referências utilizadas para o desenvolvimento deste trabalho. No Capítulo 3 é apresentada toda abordagem proposta para recomendação híbrida utilizando tags. No Capítulo 4 é descrito como foram feitos os experimentos e quais avaliações foram utilizadas para validação. No Capítulo 5 são apresentadas as conclusões obtidas e as propostas de trabalho futuro.

2 REFERENCIAL TEÓRICO

Neste capítulo serão descritos os conceitos para entendimento desse trabalho, em que são abordados os repositórios institucionais, o sistema de gerenciamento de bibliotecas utilizado, os sistemas de recomendação de conteúdo e suas variações, métricas utilizadas para avaliação das recomendações e trabalhos relacionados.

2.1 Repositórios Institucionais

Ultimamente, as iniciativas relacionadas a democratização da produção científica tem alcançado uma grande importância na pesquisa. Umas das dificuldades encontradas está relacionada a disponibilização dessas informações. Por isso, tem-se discutido sobre qual melhor meio à ser adotado para disseminar, armazenar e preservar as informações, permitindo seu acesso.

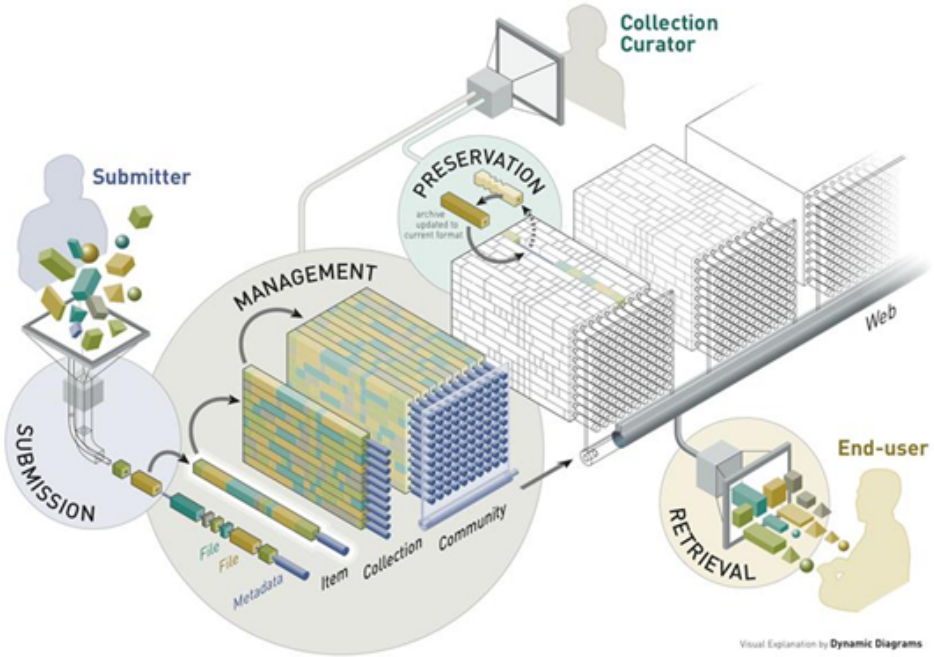
A utilização de repositórios institucionais tem sido amplamente adotada nas universidades em âmbito nacional e internacional com objetivo de compartilhar e disseminar o conhecimento produzido, promovendo o acesso à informação por um processo mais transparente e democrático, além de prover maior reconhecimento e visibilidade nacional e internacional das produções científicas das instituições.

Na Universidade Federal de Lavras – UFLA, foi implantado por meio da Biblioteca Universitária, o Repositório Institucional da UFLA (RIUFLA), que utilizou o software DSpace para desenvolvimento e gerenciamento do repositório (MEDEIROS, 2012).

O DSpace é um software sem fins lucrativos utilizado para construção de repositórios digitais abertos. Além de ser gratuito, sua instalação é simples e sua estrutura é completamente personalizável permitindo ser adaptado de acordo com a necessidade da instituição.

O software permite fácil acesso à maioria dos tipos de conteúdo digital, incluindo texto, imagens e vídeos. Possui uma comunidade de desenvolvedores comprometida com a expansão e melhoria do software. A arquitetura do DSpace é definida na Figura 2.1 (DSPACE, 2014).

Figura 2.1 – Arquitetura DSpace



Fonte: DSpace (2014).

A interface DSpace é baseada na Web e pode ser controlada de acordo com a responsabilidade de cada usuário no processo. As interfaces podem ser definidas para administradores, usuários finais, administradores do sistema e usuários envolvidos no processo de submissão de documentos. O propósito dessa separação é manter a segurança e responsabilidades bem definidas.

O sistema DSpace utiliza workflows para definir o fluxo dos processos de trabalho. Isso permite que as funções e responsabilidades de cada usuário integrante do sistema sejam bem estruturadas, mantendo uma sequência organizada de cada atividade a ser executada.

O sistema utiliza um padrão sobre o Dublin Core, esquema que define os objetos digitais (metadados), e por meio desses metadados é possível cadastrar as informações e realizar pesquisas do conteúdo na base de dados.

A arquitetura é formada por três camadas: camada de aplicação, de armazenamento e de negócios. A camada de armazenamento é implementada usando um sistema de arquivos que é gerenciado através de um banco de dados, considerado como nativo o PostgreSQL. Na camada de negócios são definidas as funcionalidades, incluindo fluxo de trabalho, gerenciamento, administração e pesquisa de conteúdo. A camada de aplicação abrange a interface de apresentação web do sistema (SMITH et al., 2003).

2.2 Sistema de Gerenciamento de Biblioteca

A evolução tecnológica vem alterando e facilitando as atividades em diversos segmentos. No caso das bibliotecas, a automatização dos processos vem se tornando um fator cada vez mais constante e necessário para melhorar a prestação de serviços, o que conduz a adoção de sistemas informatizados utilizados para dinamizar e facilitar o controle das operações.

Um sistema de gerenciamento de biblioteca assim como um repositório institucional, possui estruturas de armazenamento com dados semelhantes. Ambos armazenam metadados em sua base de dados como, título do livro/documento, nome do autor do livro/documento, palavras chave que classificam o conteúdo de um livro/documento, e podem ser utilizados para proporcionar informações úteis aos usuários.

Com base nesses dados, percebeu-se a importância das informações armazenadas no sistema PERGAMUM de gerenciamento, utilizado na Biblioteca da Universidade Federal de Lavras. O sistema é proprietário e possui acesso via browser. Pode ser utilizado com banco de dados Oracle, SQL Server ou Sybase,

possuindo alta capacidade de armazenamento, sendo sua fonte de dados o ponto principal para complementar essa pesquisa (PERGAMUM, 2015).

Este sistema armazena um histórico de todas as atividades realizadas por seus usuários. Para essa pesquisa, o ponto de maior interesse sobre essa base de dados, está relacionado ao histórico de empréstimos de livros dos usuários. Essas informações podem ser úteis para tentar identificar o perfil e o interesse de cada usuário da biblioteca.

Sendo assim, torna-se essencial que exista uma conexão ou integração entre esses sistemas para haver troca de informações para que possam ser interpretadas em busca de conhecimento. Com base nessas informações, estudos podem ser feitos para melhorar a disponibilização dos recursos para os usuários, podendo utilizar, por exemplo, recomendação de conteúdo sobre os dados disponíveis nesses sistemas.

2.3 Sistema de Recomendação

Os sistemas de recomendação (SR) são amplamente utilizados para identificar interesses das pessoas sobre um determinado produto ou serviço. É preciso conhecer os hábitos e interesses de cada indivíduo, para que seja possível recomendar um produto que realmente o consumidor deseja.

Os SR auxiliam os usuários na identificação eficaz de itens adequados a sua necessidade ou preferência, orientando-os de forma personalizada e sugerindo itens que sejam relevantes entre uma grande quantidade de opções. Com isso, evita-se a sobrecarga de informações irrelevantes melhorando o acesso à informação para usuários que não tem conhecimento detalhado sobre o item pesquisado (BARRAGáNS-MARTÍNEZ et al., 2010).

Esse tipo de sistema opera por meio de buscas sobre grandes volumes de informações que podem estar dispersas na internet ou concentrados em bases de

dados. O principal objetivo desse sistema é refinar a busca por um resultado que forneça o retorno de informações relevantes para o usuário.

Um sistema de recomendação de conteúdo pode ser classificado em dois tipos: Baseado em Conteúdo ou Colaborativo. O primeiro tem foco em um único usuário, as recomendações são relacionadas especificamente para aquele usuário de acordo com a similaridade dos itens de seu interesse. O segundo considera um grupo de usuários, e faz recomendações conforme os interesses existentes entre grupos de usuários que apresentam certa semelhança (GARCIA; FROZZA, 2013).

Complementando, Serrano-Guerrero et al. (2011) classifica um sistema de recomendação não apenas em duas, mas em três categorias principais. As classificações citadas anteriormente identificam dois tipos de classificação, baseado em conteúdo e colaborativo. Entretanto existe outra forma que é identificada como híbrida, que pode ser aplicada combinando o sistema baseado em conteúdo com o sistema colaborativo.

Uma etapa importante existente em um sistema de recomendação é o processo de coleta de informações. A coleta de informações pode ser feita de duas formas. Na forma explícita, em que as informações são absorvidas pelo sistema quando um determinado usuário informa seus gostos ou preferências e na forma implícita, utilizando-se análise de comportamento, através da captura dos caminhos (links) percorridos em um site (logs) por um usuário.

A coleta de informações explícitas está relacionada a avaliações feitas por um usuário, quais suas preferências e interesses. Essas informações também podem ser coletadas através de formulários previamente preenchidos por usuários.

A descoberta de informações implícitas é um processo mais oneroso pois, dependendo do domínio da aplicação, estas podem estar ocultas. Alguns métodos podem ser utilizados para extrair informações implícitas a partir da disponibilização de dados do usuário, como por exemplo: atividades e comportamento, relaci-

onamentos existente entre usuários, mapeamento dos itens que foram visitados, e tempo gasto na observação de um item (GARCIA; FROZZA, 2013).

2.3.1 Sistema de Recomendação Colaborativo

Um sistema de recomendação baseado em colaboração permite gerar sugestões para determinado usuário considerando opiniões e interesses similares de outros usuários, de acordo com suas características ou comportamentos que se assemelham (KARDAN; EBRAHIMI, 2013).

Existem métricas que podem ser utilizadas em técnicas de filtragem colaborativa para descobrir semelhanças entre usuários. As mais comumente usadas são cosseno (2.1), correlação de Pearson (2.2) e diferença média ao quadrado (MSD) (2.3), que são especificadas a seguir, respectivamente:

$$simicoss(x,y) = \frac{\sum_{i=1}^{n} r_{x,i} r_{y,i}}{\sum_{i=1}^{n} r_{x,i}^2 \sum_{i=1}^{n} r_{y,i}^2} \tag{2.1}$$

$$simp(x,y) = \frac{\sum_{i=1}^{n} (r_{x,i}\bar{r}_x) - (r_{y,i}\bar{r}_y)}{\sum_{i=1}^{n} (r_{x,i} - \bar{r}_x)^2 \sum_{i=1}^{n} (r_{y,i} - \bar{r}_y)^2} \tag{2.2}$$

$$simmsd(x,y) = \frac{1}{n} \sum_{i=1}^{n} (r_{x,i} - r_{y,i})^2 \tag{2.3}$$

em que $r_{x,i}$ e $r_{y,i}$ representam a média das taxas de avaliação de um usuário x e um usuário y para o item i. Com essas funções é possível calcular a similaridade entre usuários x e y. Nesse estudo adota-se a similaridade do cosseno que foi utilizada de forma satisfatória em alguns estudos como em Kim et al. (2010), Durao e Dolog (2010), Zhu et al. (2011), Oliveira e Coello (2013), Usharani e Iyakutti (2013) e Chirita et al. (2007).

Um ponto importante que deve ser levado em consideração na aplicação das métricas é a escolha da quantidade de usuários semelhantes para geração das recomendações. Uma quantidade de usuários muito pequena pode não ser sufi-

ciente para gerar regras consistentes, pois a probabilidade desses usuários terem classificado um item que um indivíduo possa ter interesse é pequena. Entretanto, uma quantidade de usuários muito elevada pode generalizar as regras, pelo fato de que haverá uma proporção maior destes usuários que não se assemelham a características de outros indivíduos. Além disso, a elevada quantidade de usuários pode comprometer a eficácia do sistema (SANCHEZ et al., 2008).

Um dos problemas relacionados à filtragem de conteúdo colaborativo, também conhecido como Partida a Frio (Cold Start Problems), é que qualquer item que seja adicionado recentemente não será recomendado até que seja avaliado por outros usuários. Outro problema, Dados Esparsos (Data Sparseness Problem), é que o número de avaliações de usuários comparado à quantidade de itens é pequeno, não sendo suficiente para chegar a uma recomendação de qualidade (KARDAN; EBRAHIMI, 2013).

2.3.2 Sistema de Recomendação Baseado em Conteúdo

O objetivo de um sistema de recomendação baseado em conteúdo é possibilitar a sugestão de itens que esteja relacionado aos interesses do usuário de acordo com seu perfil. Para que isso seja possível é necessário que o sistema identifique as semelhanças entre usuários que avaliaram determinado conteúdo. Isso possibilitaria a seleção de apenas o conteúdo que tem relação com as preferências do usuário (BARRAGáNS-MARTÍNEZ et al., 2010).

A abordagem de filtragem baseada em conteúdo classifica a preferência de usuários por itens semelhantes que anteriormente foram identificados de seu interesse. Com base nas informações relacionadas ao conteúdo e dados específicos sobre um usuário, é possível identificar se há alguma relação entre o usuário e o conteúdo (JR; OLIVEIRA, 2011).

Essa abordagem possui características semelhantes aos sistemas de filtragem e recuperação da informação. A evolução dessas técnicas teve grande impor-

tância devido a sua aplicação sobre sistemas amplamente constituídos de textos e que são muito utilizados atualmente. Isso repercutiu nos sistemas baseados em conteúdo que utilizam informações textuais, como documentos, sites e feed de notícias para gerar sugestões (ADOMAVICIUS; TUZHILIN, 2005).

Para realizar a filtragem de conteúdo pode-se utilizar o modelo espaço vetorial. Com este modelo é possível selecionar um vetor de itens que mais se assemelha a um item chave. Com base neste item chave, outros vetores devem ser selecionados. Assim, o modelo espaço vetorial seleciona um item desejado e calcula a semelhança existente entre cada um dos itens desses vetores com base no item chave. A equação desse modelo é representada abaixo em (2.4):

$$sim(u,q) = \frac{\sum_i (u(i) \times q(i))}{\sqrt{\sum_i u(i)^2} \times \sqrt{\sum_i q(i)^2}} \tag{2.4}$$

Em que, u representa os usuários e q a quantidade de itens. O cálculo é feito para todos os vetores de itens que são classificados com base no resultado do cálculo.

A filtragem por conteúdo permite recomendar conteúdo sem que este tenha sido anteriormente avaliado. Ao contrário da filtragem baseada em colaboração que depende de avaliações sobre o conteúdo para ser aplicada (BARRAGáNS-MARTÍNEZ et al., 2010).

2.3.3 Sistema de Recomendação Híbrido

Para tentar resolver as limitações dos sistemas baseados em conteúdo e dos colaborativos, desenvolveu-se o sistema híbrido, que supri as limitações das duas técnicas através da combinação desses sistemas. Dessa forma, quando não existem informações dos usuários e suas avaliações, aplicam-se as técnicas usadas em sistemas baseado em conteúdo. Caso não existam informações suficientes sobre o conteúdo associado aos itens, então são aplicadas as técnicas usadas nos sistemas colaborativos (KARDAN; EBRAHIMI, 2013).

Existem várias abordagens que utilizam sistemas híbridos. Uma delas é apresentada por Castro-Herrera (2010), em que o objetivo principal é encontrar usuários relevantes em fóruns de acordo com suas postagens. A filtragem colaborativa é utilizada para calcular a similaridade e identificar os usuários que contribuíram com mensagens, apresentando os mesmos interesses. Por meio da filtragem de conteúdo, as palavras chaves são identificadas e a contagem da frequência é realizada atribuindo pesos sobre cada ocorrência da palavra. Dessa forma, é gerado um resultado com a aplicação de diferentes conceitos através da fusão de técnicas.

Outra abordagem de sistema híbrido é apresentada por Salter e Antonopoulos (2006), que utiliza os resultados obtidos pela filtragem colaborativa como dados de entrada para a filtragem baseada em conteúdo. Nesse sistema, encontram-se primeiramente os usuários com maior proximidade de acordo com as avaliações feitas sobre filmes. Com base nas características (metadados) do filme e dos usuários, e também de novos filmes ainda não avaliados, calcula-se novamente a relação de proximidade.

2.3.4 Sistema de Recomendação Utilizando Tags

A utilização de etiquetas (tags) vem sendo adotado como uma alternativa para facilitar a classificação de conteúdo com base no senso comum dos usuários. Utilizando tags, um usuário é capaz de associar um item a determinando conteúdo com propósito de descrevê-lo utilizando uma palavra para classificá-lo conforme a informação transmitida.

Em alguns casos, como na web, a classificação utilizando tags pode ser muito útil devido a grande quantidade de itens existentes para classificação.Não havendo uma pessoa específica para realizar essa tarefa, os próprios usuários podem ser categorizadores daquele conteúdo, ou seja, qualquer usuário pode fazer anotações sobre um conteúdo utilizando qualquer tipo de tag. A esse tipo de prática é dado o nome de etiquetagem colaborativa (OLIVEIRA; COELLO, 2013).

O uso de tags pode ser uma alternativa para melhorar o desempenho dos sistemas de recomendação, assim passa a ser considerada a relação entre três elementos, usuários, tags e itens. Conforme citado anteriormente, a recomendação colaborativa apresenta o problema de dados esparsos e partida a frio, ambas as limitações podem ser amenizadas utilizando tags. Sendo assim, aplica-se a abordagem de etiquetagem colaborativa, em que vários usuários classificam determinado conteúdo utilizando palavras chaves (tags), o que possibilita identificar preferências de acordo com as marcações (KIM et al., 2010).

Existem diferentes tipos de sistemas de marcações de tags, um deles está relacionado a marcação por comportamento, que pode ser classificado em auto marcação, baseado em permissão e marcação livre. Em auto marcação, os usuários marcam apenas conteúdo criado por eles mesmos, para que possa recuperar futuramente. Na marcação baseada em permissão, os usuários possuem diferentes níveis de permissão para realizar tais marcações. Em marcação livre, qualquer usuário pode marcar vários itens.

Outros sistemas de marcação, que levam em consideração a agregação de tags, são conhecidos como bag-model e set-model. O set-model não permite a repetição de tags para marcação de um item, assim o usuário não possui muita liberdade para realizar a marcação, já que o sistema disponibiliza apenas algumas tags para classificação de determinado item. Já o bag-model permite que tags duplicadas possam ser utilizadas por diferentes usuários para classificar o mesmo item (MARLOW et al., 2006).

2.4 Cálculo de Peso das Tags

A técnica TF-IDF (Term Frequency – Inverse Document Frequency) é utilizada em mineração de texto para identificar a importância das palavras em um texto. Um valor é atribuído para cada termo extraído do texto de acordo com a frequência da palavra no texto ou nos documentos. Esse valor é utilizado para

representar o peso que determina a importância do termo no texto e em todos os documentos da base de dados. Na equação (2.5) é representado como é feito o calculo do TF (ELMASRI; NAVATHE, 2011):

$$TF_{ij} = \frac{f_{ij}}{\sum_{i=1}^{|V|} f_{ij}} \tag{2.5}$$

Nessa equação os símbolos tem o seguinte significado:

- TF_{ij} é a frequência do termo normalizada do termo i no documento D_j.
- f_{ij} é o número de ocorrências do termo i no documento D_j.

O cálculo do IDF é feito como apresentado na equação (2.6):

$$IDF_i = \log(\frac{N}{n_i}) \tag{2.6}$$

Onde:

- IDF_i é o peso de frequência do documento inverso para o termo i.
- N é o número de documentos na coleção.
- n_i é o número de documentos em que o termo i ocorre.

O TF-IDF utiliza o produto da frequência normalizada de um termo $i(TF_{ij})$ no documento D_i e a frequência inversa do documento do termo i para determinar o peso de um termo no documento. O cálculo é representado na equação (2.7):

$$TF - IDF_{ij} = TF_{ij} \times IDF_{ij} \tag{2.7}$$

A utilização apenas da frequência pode induzir a erro, pois nem sempre as palavras que são mais frequentes são as mais importantes. Por isso, calcular o peso com TF-IDF pode ser útil para determinar um valor mais confiável de relevância do termo no documento.

Esse cálculo foi adaptado para ser aplicado sobre os dados, em que considera-se que um livro emprestado por um usuário é um documento e os termos são as informações extraídas dos seus metadados, como por exemplo, título, palavras chave e autor. Assim, vários livros representam um conjunto de documentos. Para o calculo do TF-IDF considera-se que documentos equivalem a livros e os seus metadados equivalem a termos que fazem parte de um documento.

$$TF_{tl} = \frac{f_{tl}}{\sum_{i=1}^{|V|} f_{tl}} \tag{2.8}$$

onde:

- TF_{tl} é a frequência do termo normalizada do termo t no livro L_i.
- f_{tl} é o número de ocorrências do termo t no livro L_i.

O cálculo do IDF é feito como apresentado na equação (2.6):

Para o cálculo do IDF representado na equação (2.9) considera-se:

$$IDF_t = \log(\frac{N}{n_t}) \tag{2.9}$$

Onde:

- IDF_t é o inverso do peso de frequência do livro para o termo t.
- N é a quantidade de livros do conjunto.
- n_t é o número de livros em que o termo t ocorre.

Assim temos o calculo do TF-IDF na equação (2.10):

$$TF - IDF_{tl} = TF_{tl} \times IDF_{tl} \tag{2.10}$$

Descrevendo a adaptação da fórmula, tem-se que o TF é a fração entre a quantidade de vezes que um termo ocorre no conjunto de livros emprestados, pelo

número total de termos do conjunto de livros. E o IDF é a fração da quantidade total de livros pela quantidade de livros que um termo ocorre. Considere que o usuário Miguel emprestou os livros “Programação Java”, “Introdução ao C++” e “Redes de Computadores”. Extraímos os metadados desses livros e obtemos as seguintes informações para os livros, conforme tabela 1:

Tabela 2.1 – Books and terms (tags)

Livros	**Termos**		
Programação Java	programação	orientada	objetos
Introdução ao C++	programação	objetos	C++
Redes de Computadores	tcpip	camadas	segurança

Aplica-se o cálculo do TF-IDF e obtém-se o peso de cada termo. Para o termo “objetos” tem-se o seguinte resultado:

$$TF_{tl} = \frac{2}{9} = 0.222$$
$$IDF_t = \log(\frac{3}{2}) = 1,5$$
$$TF - IDF_{tl} = 0,222 \times 1,5 = 0,333$$

No calculo de TF, o valor 2 representa a quantidade de ocorrências do termo “objetos” em todo conjunto de termos. O valor 9 representa a quantidade total de termos do conjunto. No calculo do IDF, o valor 3 representa a quantidade de livros do conjunto e o valor 2 a quantidade de livros em que o termo “objetos” ocorre. Como resultado do cálculo TF-IDF, obtém-se para o termo “objetos” o valor de peso 0,333.

Adaptou-se cálculo, pois geralmente determina-se a importância do termo no documento, mas nesse caso, a importância do termo é identificada para um determinado usuário. Feito isso, a lista de tags do usuário é preenchida com os pesos de relevância encontrados para cada termo.

2.4.1 Avaliação das Recomendações

Em sistemas de recomendação, as métricas de precisão, revocação e F-Score, são utilizadas para avaliar se o sistema recomenda itens que realmente sejam considerados relevantes pelo usuário. Para possibilitar o calculo dessas medidas é necessário quantificar e categorizar os itens quanto as suas informações (TEJEDA-LORENTE et al., 2014). Os itens podem ser classificados em relevante ou irrelevante e recomendado ou não recomendado, conforme tabela 2.2:

Tabela 2.2 – Tabela de relações

	Recomendado	**Não recomendado**
Relevante	$N_{rr}(VP)$	$N_{rn}(FN)$
Irrelevante	$N_{ir}(FP)$	$N_{in}(VN)$

Essa tabela utiliza a mesma concepção da matriz de confusão, em que termos como verdadeiro positivo (VP), falso positivo (FP), falso negativo (FN) e verdadeiro negativo (VN) são aplicados para comparar determinada classificação de um item com a classificação correta desejada, e seus valores são utilizados nos cálculos de precisão, revocação e f-score.

Precisão é definida pela razão entre os itens relevantes recomendados por itens recomendados. Essa métrica é utilizada com objetivo de medir a probabilidade de um produto recomendado ser relevante para o usuário. A métrica é definida na equação (2.11):

$$P = \frac{N_{rr}}{N_{rr} + N_{ir}} \tag{2.11}$$

Revocação é calculada pela razão entre itens relevantes recomendados por itens relevantes. Representa a probabilidade de um item relevante ser recomendado. A métrica é definida na equação (2.12):

$$R = \frac{N_{rr}}{N_{rr} + N_{rn}} \tag{2.12}$$

F-Score (F) é a média harmônica da combinação dos valores de precisão e revocação. A medida é utilizada para comparar diferentes conjuntos de resultados. A métrica é definida na equação (2.13):

$$F = \frac{2 \times R \times P}{R + P} \tag{2.13}$$

2.4.2 Trabalhos Relacionados

Os sistemas de recomendação auxiliam os usuários a identificar de forma eficaz itens de acordo com seu interesse e necessidade. Dentro de uma grande variedade de opções, esse sistema orienta os usuários de forma personalizada para acessar conteúdo que seja relevante e útil. Esse tipo de sistema é utilizado para reduzir a sobrecarga de informação como em Tejeda-Lorente et al. (2014), sendo aplicado em diversos meios, como nos sites de e-commerce para melhorar vendas, em sistemas para indicação de filmes e programas de tv, em blogs para recomendação de artigos, nas redes e mídias sociais (DURAO; DOLOG, 2012), (BARRAGáNS-MARTÍNEZ et al., 2010), (GARCIA; FROZZA, 2013) e (LIU et al., 2015).

No trabalho de Tejeda-Lorente et al. (2014), propõe-se um sistema de recomendação baseado na qualidade dos recursos. O sistema utiliza a qualidade dos artigos para estimar sua relevância e aplica uma abordagem baseada em lógica fuzzy, em que os usuários fornecem suas preferências por meio do conceito da linguística fuzzy para montar o seu perfil. Essa estratégia aplica uma abordagem simples baseada em conteúdo, mas prevê a utilização de uma aplicação de recomendação híbrida ao adaptar a filtragem colaborativa a partir da experiência de recomendações compartilhadas entre os usuários. Além disso, um módulo de reclassificação do conteúdo é incorporado, que combina a relevância estimada de um item com sua qualidade. Considera-se que os recursos com maior preferên-

cia pelos usuários são os que possuem boa qualidade. Com isso, tende-se a gerar recomendações mais úteis e precisas.

O estudo feito por Lai, Liu e Lin (2013), acrescenta a confiabilidade dos usuários na técnica de filtragem colaborativa para tentar melhorar a qualidade da recomendação e desenvolve um modelo de confiança híbrido. O sistema calcula o valor da confiança de acordo com avaliações dos usuários sobre itens já classificados. Nesse caso, um usuário é considerado de alta confiança quando contribui com classificações mais precisas do que outros. Os usuários que possuem alguma semelhança são identificados e assim formam-se grupos para compartilhar itens entre os seus integrantes.

Com isso, as preferências dos usuários afetam o grupo, aumentando a confiança pessoal, o que pode alterar as recomendações a partir da perspectiva do grupo. Em contrapartida, os valores de confiança podem não ter uma precisão adequada quando a quantidade de itens classificado por um usuário é baixa. Por isso, torna-se necessário a utilização de um modelo de confiança híbrido, que considere as classificações pessoais e de todo o grupo. Utilizou-se a abordagem tf-idf para determinar a importância dos documentos.

Um documento que o usuário faz o download ou upload é considerado mais importante do que aquele que foi visitado. Assim, de acordo com os acessos, o perfil dos usuários vai sendo montado. A partir desse momento, a semelhança entre os usuários é calculada por meio do cosseno e os grupos são formados com base em seus valores de proximidade. Os documentos que possuírem um alto índice de previsão são indicados em uma lista de recomendação. Esse método ainda melhora o processo de previsão, pois utiliza a similaridade entre os usuários para fazer recomendações.

A pesquisa feita por Moreno et al. (2016), propõe uma estrutura completa para lidar com os problemas mais presentes em sistemas de recomendação: escalabilidade, dispersão e arranque a frio. Neste trabalho aborda-se o contexto de

recomendação de filmes, mas pode ser aplicado em outros domínios. A proposta consiste em combinar métodos de mineração web e ontologias para tentar induzir modelos em dois níveis de abstração.

O modelo de nível mais baixo é construído a partir de dados que não utilizam informações semânticas, enquanto que o modelo de mais alto nível utiliza dados da web classificados com informações semânticas de acordo com a ontologia definida. Com a combinação desses modelos, são gerados padrões com alto nível de abstração, capaz de relacionar tipos de produtos e perfis de usuários de modo generalizado. Assim, o modelo é capaz de recomendar produtos que ainda não foram avaliados ou fazer recomendações para novos usuários. Além disso, o modelo desenvolvido é capaz de lidar com o problema de dispersão utilizando métodos de classificação associativa.

Na literatura também existem trabalhos que abordam a utilização de tags para melhorar a qualidade das recomendações, como é o caso do estudo científico desenvolvido por Kim et al. (2010), que utiliza filtragem colaborativa combinada com etiquetagem colaborativa para aprimorar recomendações usando tags criadas por usuários. Nessa pesquisa, a abordagem conhecida como marcação colaborativa gera perfis de usuários conforme as marcações feitas nos itens. Isso é feito em duas fases: uma fase de construção de modelos e outra de recomendação probabilística.

Usando um esquema de filtragem colaborativa, são geradas as marcações (conjunto de tags candidatas), que poderão ser direcionadas para um usuário. Por meio dessas tags, o algoritmo Naive Bayes é aplicado para recomendar os itens. O sistema de marcação é composto pela relação de três elementos: usuário, item e termo. Estes três elementos são analisados considerando sua frequência sobre um vetor binário, por meio da relação usuário-item, usuário-tag e tag-item, em que o usuário faz uma marcação em um item utilizando algum termo. Por meio do processo de filtragem colaborativa o sistema encontra outras marcações semelhantes

às tags do conjunto de tags candidatas e as recomenda para os usuários de acordo com a similaridade das tags pessoais de cada usuário.

Os dados utilizados nessa pesquisa foram coletados no site del.icio.us, que é um serviço de marcação social que utiliza marcação colaborativa. A pesquisa identificou que o algoritmo proposto obteve resultados significativos em relação a qualidade das recomendações, além de ter diminuído os problemas relacionados a dados esparsos e arranque a frio. Observou-se também que o método forneceu itens mais adequados ao perfil do usuário, mesmo que a quantidade de itens recomendados fosse menor. Foram identificados problemas relacionados a polissemia e sinonímia, que afeta a qualidade do conjunto de tags, mas é uma limitação que pode ser amenizada futuramente com aplicação de semântica.

Outras pesquisas científicas utilizando palavras chaves (tags) são adotadas em outros estudos aplicados a recomendação de conteúdo. Uma série de artigos foram desenvolvidos por Durao e Dolog (2012), Durao e Dolog (2010), Durao e Dolog (2009), Dolog et al. (2011) e Durao e Dolog (2014), propondo melhorar a qualidade das recomendações em cada etapa da pesquisa por meio da combinação de métodos aplicada a recomendação de conteúdo híbrida, tags e filtragem colaborativa. Nessas pesquisas, páginas de hipertexto como wikis são vasculhadas a procura de tags (marcações/palavras chaves) que possam representar algum conteúdo, ou seja, que identificam e transmitem algum significado sobre a página. Com base no acesso dos usuários aos links de uma página que o redireciona a outra, as escolhas vão sendo contabilizadas de acordo com a palavra que serviu como âncora para acesso a nova página. As marcações feitas pelos usuários são uma forma de classificar e avaliar a página, e torna possível a criação do perfil que permite o direcionamento de sugestões a serem inseridas na página.

É perceptível que a maioria dos estudos consideram as limitações apresentadas pelas técnicas de recomendação, e por isso adotam uma abordagem híbrida

e incrementam outros recursos, como a formação de tags para combinar mais de uma técnica e tentar suprir essas limitações.

3 ABORDAGEM PROPOSTA PARA RECOMENDAÇÃO HÍBRIDA BASEADA EM TAGS

3.1 Metodologia e Processo de Trabalho

O estudo para recomendação de conteúdo utilizando tags iniciou-se com a extração dos metadados dos livros que foram emprestados pelos usuários da biblioteca. Os metadados foram extraídos da base de dados da biblioteca sendo pré-processados e transformados em tags que formam uma lista de tags do usuário. Nesse estudo considera-se que o interesse do usuário está relacionado com os livros que eles emprestaram. Com isso, torna-se possível identificar por meio das tags extraídas dos metadados dos livros quais assuntos são de seu interesse.

Com base no histórico de empréstimos filtragem de conteúdo é aplicada para criar o modelo espaço vetorial de itens. Com este modelo seleciona-se um vetor de itens (livros) que mais se assemelha a um item chave. Ao selecionar um item desejado, a semelhança existente entre cada um dos itens desses vetores é calculada com base no item chave (BARRAGáNS-MARTÍNEZ et al., 2010). Nesse modelo, aplica-se alguns conceitos dessa técnica para coletar informações que sejam importantes para designar o interesse do usuário.

Os dados textuais dos metadados dos livros são extraídos para gerar as tags e utiliza-se a medida estatística TF-IDF (Term Frequency – Inverse Document Frequency) para estabelecer os pesos de importância para cada tag. O cálculo do TF-IDF foi adaptado para esse estudo, sendo assim considera-se que documentos equivalem a livros e os seus metadados equivalem a termos que fazem parte de um documento.

Com as listas de tags definidas, aplica-se a filtragem colaborativa e a partir do calculo da similaridade do cosseno combinado com o calculo do TF-IDF identifica-se os usuários que possuem preferências semelhantes de acordo com o peso de cada tag do usuário. Com isso, livros que são emprestados por um usuário,

mas que podem interessar a outro de acordo com um grau de semelhança são indicados. Nessa aplicação os usuários que possuem um grau de similaridade maior ou igual a 95% recebem como recomendação os livros que foram emprestados por eles. Considera-se um valor percentual maior para evitar recomendações entre usuários que tenham pouca relação de semelhança, diminuindo a probabilidade de sugerir livros que não sejam adequados ao perfil do usuário.

Com base nas listas de tags, os usuários recebem como recomendação uma lista de livros e documentos que devem ser avaliadas. De acordo com as avaliações retornadas, as listas de tags se modificam conforme o interesse expressado pelo usuário nas avaliações. As medidas de precisão, revocação e f-measure para avaliar a exatidão das recomendações.

3.2 Estrutura do Sistema

A implementação do sistema foi desenvolvida utilizando tecnologias open-source. A aplicação principal, responsável pelo tratamento dos dados e por todas as operações necessárias que geram as recomendações para os usuários foi implementada utilizando linguagem de programação Java.

Os dados coletados foram obtidos por meio de conexão a dois bancos de dados diferentes: uma conexão é feita ao banco de dados SQL Server, que armazena as informações dos registros de empréstimos dos usuários da biblioteca universitária, além de todos os metadados de livros e demais informações gerenciadas pelo sistema Pergamum. Outra conexão é realizada no banco de dados PostgreSQL do Repositório Institucional para coletar informações dos metadados provenientes dos documentos digitais.

Os dados que foram coletados das duas bases são pré-processados e aplicados ao algoritmo, que gera as recomendações e as armazena em um banco de dados MySQL. As informações de recomendação são exibidas em uma aplicação WEB, que disponibiliza aos usuários uma lista de recomendações que devem ser

avaliadas com uma nota conforme seu interesse. Para ter acesso as recomendações, os usuários recebem em seu e-mail um link que irá direcioná-los para a página web que contém sua lista de recomendações. Essa aplicação web foi desenvolvida utilizando linguagem PHP, JavaScript, HTML e BootStrap.

Para exibir os metadados na lista de recomendações da página web de cada usuário, utilizou-se um WebService que retorna os dados em formato XML, que em seguida são convertidos pelo PHP para o formato JSON. Um código JavaScript faz a leitura do JSON e envia os dados para serem exibidos no navegador. O processo para inserir as informações dos documentos do repositório institucional é semelhante, porém as informações são retornadas via SOLR e não por um XML.

As avaliações (feedback) dos usuários sobre a lista de recomendação são armazenadas na base de dados MySQL. Essas informações são utilizadas para remodelar o perfil dos usuários e gerar novas recomendações.

Para otimizar a inserção e consulta dos dados pela aplicação, o banco de dados MySQL foi substituído pelo MongoDB, que é um banco de dados NoSQL orientado a documentos. A utilização do MongoDB proporcionou uma redução considerável do tempo de inserção dos dados em relação ao MySQL. Porém, utilizou-se o MySQL inicialmente para facilitar a coleta e análise de informações para a pesquisa.

A implementação e execução dos experimentos foi realizada em um computador Intel (R) Core (TM) i5-3470 CPU 3.20 GHz com 4 GB de RAM e sistema operacional Windows 7.

3.3 Coleta de Informações e Feedback

Uma etapa importante para formação do perfil de um usuário em um sistema de recomendação é o processo de coleta de informações, que geralmente pode ser feita de duas formas. Na forma explícita, em que as informações são absorvidas pelo sistema quando um determinado usuário informa suas preferências,

e na forma implícita, ao utilizar análise de comportamento, por meio da captura dos caminhos (links) percorridos por um usuário em um site.

A coleta de informações explícitas é aplicada quando usuários fazem avaliações e indicam suas preferências e interesses, que podem ser coletadas por meio preenchimento de formulários. A descoberta de informações implícitas é um processo mais oneroso, pois dependendo do domínio da aplicação as informações podem estar ocultas. Alguns métodos são utilizados para extrair informações implícitas a partir da disponibilização de dados do usuário, como por exemplo: atividades e comportamento, relacionamentos existente entre usuários, mapeamento dos itens que foram visitados, e tempo gasto na observação de um item (GARCIA; FROZZA, 2013).

Após as primeiras recomendações, torna-se importante saber a opinião (feedback) dos usuários sobre a relevância do que foi sugerido para aprimorar a acurácia do sistema em futuras indicações de conteúdo. O feedback pode ser fornecido de forma implícita e explícita. Na forma explícita os usuários retornam suas opiniões por meio de avaliações (ratings), comentários textuais ou informam em escala binária se tem ou não interesse no conteúdo. Quando é aplicado o feedback implícito, técnicas automáticas de inferência monitoram as ações dos usuários para descobrir suas preferências. Para que isso seja possível, utiliza-se o histórico de navegação do usuário, links que são acessados e tempo consumido em uma página.

A combinação dos dois métodos permite avaliar se o comportamento do usuário identificado na forma implícita está de acordo com suas avaliações recuperadas explicitamente. Assim, constrói-se uma alternativa híbrida capaz de verificar se os atos do usuário seguem uma lógica, ou seja, se estão condizentes dentro de seus padrões de interesses (REATEGUI; CAZELLA; OSóRIO, 2006).

3.4 Pré-Processsamento dos Daddos

Os dados provenientes dessa pesquisa encontram-se armazenados em duas bases de dados distintas, uma refere-se ao sistema de gestão da biblioteca e outra que possui documentos do repositório institucional. No repositório institucional são armazenados documentos digitais como teses, dissertações e artigos. A base de dados da biblioteca armazena toda informação necessária para gerenciamento das atividades acadêmicas, sendo povoada diariamente por meio de ações dos usuários (alunos, professores e funcionários). Quando um usuário faz o empréstimo de um livro, várias informações sobre essa ação são armazenadas na base de dados. Essas informações contem metadados que foram utilizados para extrair tags e traçar o perfil do usuário. Os metadados são formados pelos seguintes atributos: Metadados do histórico de empréstimos de usuários:

- Nome usuário: nome do usuário que efetuou empréstimo de livro;
- Matrícula: valor numérico único que identifica um aluno no sistema;
- Cod Departamento: código do departamento ao qual o aluno que efetuou empréstimo esteja vinculado;
- Palavra Chave (tags): esse metadado está relacionado a informações contidas no livro que foi emprestado pelo usuário. Essas informações são manipuladas e transformadas em tags.

Metadados de livros:

- Título do Livro: corresponde ao nome do livro;
- Palavras chave (tags): esse campo armazena informações como autor do livro, área e palavras chaves que caracterizam o conteúdo do livro;
- Cod Acervo: valor numérico único que identifica um determinado livro;

- Classificação: código que classifica um livro de acordo com a área de concentração;
- Data de empréstimo: data referente ao empréstimo do livro feito pelo usuário.

Essas informações são utilizadas para gerar a lista de tags que forma o perfil de cada usuário. Uma lista de livros é enviada como recomendação ao usuário, que avalia com uma nota cada item que lhe foi recomendado. Assim, torna-se possível identificar com maior precisão as preferências do usuário, a fim de melhorar a qualidade das recomendações. As informações provenientes dos documentos da base de dados do repositório são integradas com a base de dados da biblioteca para gerar recomendações provenientes de ambas as bases de dados. Além de ser uma alternativa para complementar a variedade de informações, também contribui para fomentar a utilização do repositório, visto que os usuários terão por meio de uma única interface acesso a conteúdo de ambos os sistemas. Os atributos utilizados para extrair as informações dos metadados dos documentos digitais são:

- Título do documento: título do documento armazenado no repositório;
- Palavras chave: termos que são usados como palavras chave do documento;
- Resumo: descrição textual utilizada para descrever o tema tratado no documento;
- Autor: identifica as pessoas que desenvolveram o trabalho.

Nesta etapa, os dados que foram coletados da base de dados serão filtrados para manter apenas o que é relevante para a pesquisa. Para isso, são aplicadas técnicas de seleção e transformação de dados utilizada na maioria dos sistemas que trabalham com mineração de dados ou recuperação da informação.

Primeiramente as stopwords são removidas, pois mesmo que utilizadas para formação de uma sentença, raramente contribuem para agregar significado às

palavras. As stopwords são consideradas como palavras que ocorrem com uma alta frequência nos documentos, mas tornam-se potencialmente inúteis por serem muito comuns, não contribuindo para relevância na pesquisa. Exemplos dessas palavras são: o, a, e, para, onde, que, estava, isso, entre várias outras. Um arquivo de stopwords é formado para ser utilizado pelo sistema para remover das informações recuperadas todas as palavras existentes nessa lista (ELMASRI; NAVATHE, 2011).

Outras remoções são feitas para manter a qualidade das informações recuperadas. Palavras com menos de dois caracteres são desconsideradas, e para facilitar a análise são removidos acentuações, caracteres especiais e todas as palavras são convertidas para letras minúsculas.

Nos primeiros testes aplica-se o processo de stemização, também conhecido como técnica de stemming. Esse processo é utilizado em diversas aplicações computacionais para transformar as formas variantes de uma palavra em uma representação mais precisa, e que seja genérica o suficiente para capturar a essência das palavras (ALVARES, 2014). A abordagem utilizada no processo de stemming aplicado nesse estudo foi com base no algoritmo Snowball, uma variação do algoritmo de Porter, que utiliza a remoção de sufixos (XAVIER; SILVA; GOMES, 2013) e (WILLETT, 2006).

No entanto, a aplicação dessa técnica não foi satisfatória para este experimento. Ao utilizar stemming a precisão do algoritmo diminui. Uma possível causa pode estar relacionada a criação das palavras chave (tags). Nesse contexto as palavras não são criadas livremente pelo usuário, mas sim por especialistas que analisam os termos antes de cadastrar as informações nos metadados do documento. Sendo assim, considerando a perda de precisão e também a demanda por processamento computacional em executar essa tarefa, a utilização do stemming nessa aplicação não foi continuada.

3.5 Arquitetura do Sistema

Alguns trabalhos encontrados na literatura utilizam uma abordagem híbrida, combinando filtragem colaborativa e filtragem baseada em conteúdo. Em nosso trabalho a abordagem híbrida é aplicada e aprimorada para utilização de tags. Além disso, adota-se o conceito de detecção de novidade para identificar ao longo do tempo alterações de interesse dos usuários.

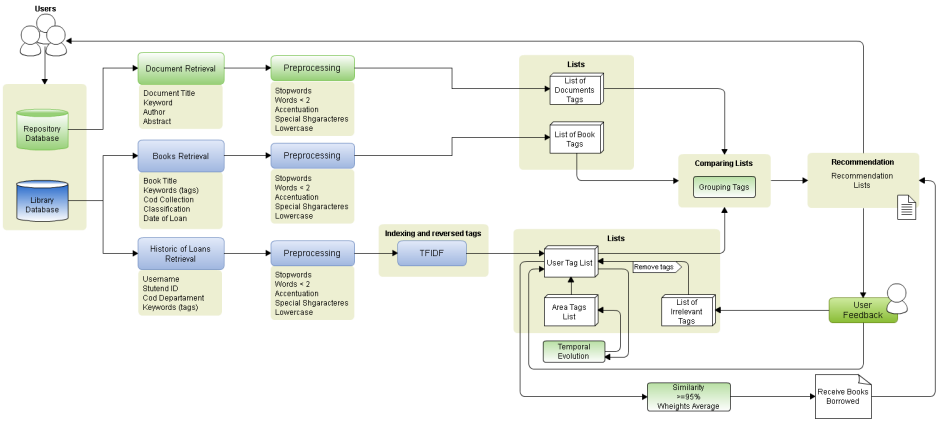
O perfil do usuário é formado pelas tags que são recuperadas dos metadados dos livros emprestados pelos usuários, assim cada usuário terá uma lista de tags. O conceito de filtragem baseada em conteúdo é aplicado para obter o histórico de empréstimos de livros dos usuários da biblioteca, que é combinado com uma lista de recomendações de livros enviada ao usuário para avaliar cada item com uma nota.

A filtragem colaborativa é utilizada para identificar a proximidade entre os perfis dos usuários e recomendar conteúdo entre aqueles que possuam alta similaridade de interesses. O conceito de detecção de novidade é aplicado para identificar as mudanças na lista de tags de cada período de um determinado curso. Na Figura 3.1, é apresentado com detalhes a estrutura do sistema e suas etapas, que serão detalhadas nos próximos tópicos.

3.6 Lista de Tags

Alguns estudos como em Jiang et al. (2010), Luo, Wei e Lai (2011) e Lin et al. (2016) adotam um modelo para representar características em documentos, conhecido na área de recuperação da informação como Bag of words, e tem sido considerado promissor na classificação de conteúdo. Neste modelo, qualquer conteúdo textual é transformado em um conjunto de palavras que transmitem algum significado e são utilizadas como base para classificação de determinado recurso.

Figura 3.1 – Arquitetura do sistema

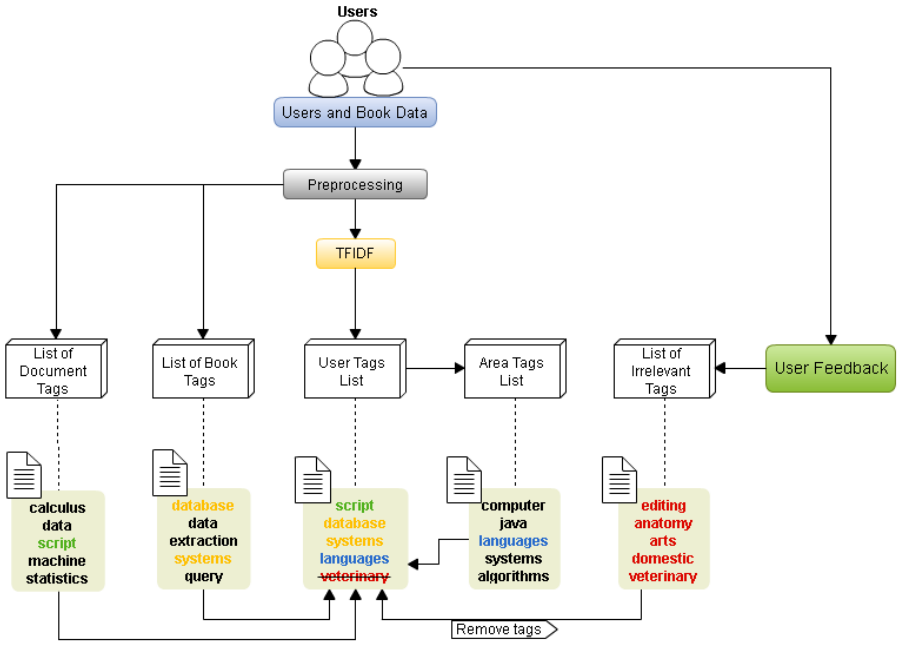


Seguindo este conceito, o modelo é adaptado para o nosso estudo para utilizar as listas de tags (bag of tags).

As listas de tags são formadas por um conjunto de palavras extraídas dos metadados de livros e são coletadas de duas formas: registro dos livros que foram emprestados por usuários na base de dados e lista (feedback) de livros recomendados para o usuário. Cada termo ou palavra é considerada uma tag que expressa um significado que classifica o conteúdo ao qual esta tag foi extraída. Nesse caso, se um usuário faz empréstimo de um livro cujo título seja “Programação de Computadores”, as tags extraídas dos metadados desse livro representam seu contexto, e consequentemente indicam que o usuário tem interesse em documentos referentes a esse assunto. É com base no conjunto de tags de cada usuário que seu perfil é traçado para recomendar conteúdo. Nesse trabalho, implementa-se um método que é constituído por cinco listas de tags, que são aplicadas durante o processo de recomendação. Essa fase está ilustrada na figura 3.1, mas com intuito de facilitar a compreensão é representada com mais detalhes na figura 3.2.

As cinco listas de tags representadas na figura 3.2 são: lista de tags de documentos, lista de tags de livros, lista de tags de usuário, lista de tags da área e lista de tags irrelevantes. Com exceção da lista de tags de documentos e livros, as

Figura 3.2 – Processo de formação das listas de tags



outras três podem ser consideradas listas de tags dinâmicas, pois são alteradas ao longo do tempo de acordo com as mudanças de interesse dos usuários.

3.6.1 Lista de Tags de Livros

Essa lista é formada a partir da extração dos metadados que classificam um determinado livro. Por meio de uma conexão ao banco de dados da biblioteca as informações dos livros contidas na base de dados do sistema são capturadas. Uma consulta retorna os metadados do livro que são extraídos e tratados na fase de pré-processamento, formando as listas de tags de livros. Essa lista é imutável, ou seja, suas tags serão sempre as mesmas para cada livro, a menos que seus metadados sejam alterados no sistema. Essa lista é representada em um vetor binário de itens x termos (i,t), onde i se refere ao item, que neste caso é o livro recuperado, e t

refere-se ao termo, que é a palavra extraída dos metadados, a qual denomina-se de tag. A Figura 3.3 ilustra como é a estrutura dessa lista.

Figura 3.3 – Vetor de livros x tags

termos (tags)

itens(livros)

	java	computing	software	web	security	programing
Introduction of language	1	1	1	0	0	1
Introduction of Computer Science	1	1	1	1	0	1
Cloud computing	1	1	1	1	1	0
Computer Network	0	1	1	1	1	0

O vetor que representa a relação itens-termos (livros-tags) é formado pela verificação da existência ou não de determinada tag em um livro. Se o livro possui a tag então recebe valor 1, caso contrário o valor será 0.

3.6.2 Lista de Documentos

A formação dessa lista segue o mesmo princípio da lista de tags de livros. No entanto, os metadados são extraídos dos documentos armazenados no repositório institucional. Uma consulta é utilizada para retornar os metadados do documento que são extraídos e tratados na fase de pré-processamento, formando-se as listas de tags de documentos. Essa lista também é imutável, ou seja, suas tags serão sempre as mesmas para cada documento, a não ser que seus metadados sejam alterados.

3.6.3 Lista de Tags de Usuário

A lista de tags de usuário é formada pelos termos extraídos dos metadados dos livros contidos na base de dados da biblioteca, considerando o histórico de empréstimo do usuário em um determinado período de tempo. Essa lista interage com a lista de tags irrelevantes e tags de área, que vão sofrendo alterações ao

longo do tempo e adaptando as tags de acordo com o comportamento do usuário no sistema. Uma lista com esse fluxo pode ser observado na Figura 3.2.

Após essa etapa, as tags são geradas e para cada uma delas é definido um peso. As tags que possuem maior valor de peso são consideradas as mais importantes na lista do usuário, pois representam alto interesse do usuário pelo conteúdo expressado pela tag. O método estatístico TF-IDF foi adaptado nesse estudo para atribuir peso as tags. Esse método é utilizado para descobrir a importância das palavras em texto não estruturado ou semi estruturado.

A lista de tags do usuário é representada em um vetor binário de usuários e termos (u,t), onde u refere-se ao usuário, que neste caso é a pessoa que realiza operações de empréstimos na base de dados, e t refere-se ao termo, que é a palavra extraída dos metadados dos livros que foram emprestados.

Figura 3.4 – Vetor de usuários x tags

termos (tags)

usuários	java	computing	software	web	security	programing
João	0,5	0,3	0,7		0,6	0,8
Pedro	0,4	0		0,4	0,5	0,7
Maria	0,8	0,6	0,6	0,7		0,8
Julia	0	0,4	0,7	0,4	0,9	0,3

No exemplo da Figura 3.4, cada tag de um determinado usuário possui um peso. Um alto valor de peso representa maior grau de importância daquele termo para o usuário. O usuário João possui as tags software e programming com maior peso, o que significa que este usuário tem mais interesse em assuntos relacionados a conteúdos que o significado da tag expressa. Os termos do vetor que não possuem valores, não fazem parte da lista de tags do usuário. Isso pode ocorrer caso os livros emprestados pelo usuário não possuam esses termos.

3.6.4 Lista de Tags Irrelevantes

Sistemas projetados para recomendar conteúdo, assim como outros tipos de sistemas, sempre estão sujeitos a erros, pois nem sempre um item que é recomendado para o usuário está de acordo com seu interesse. Quando um usuário não tem interesse em um item recomendado, classifica-se esse item como um Falso Positivo, ou seja, que foi sugerido, mas classificado como não sendo de seu interesse, e por isso não deveria ter sido recomendado.

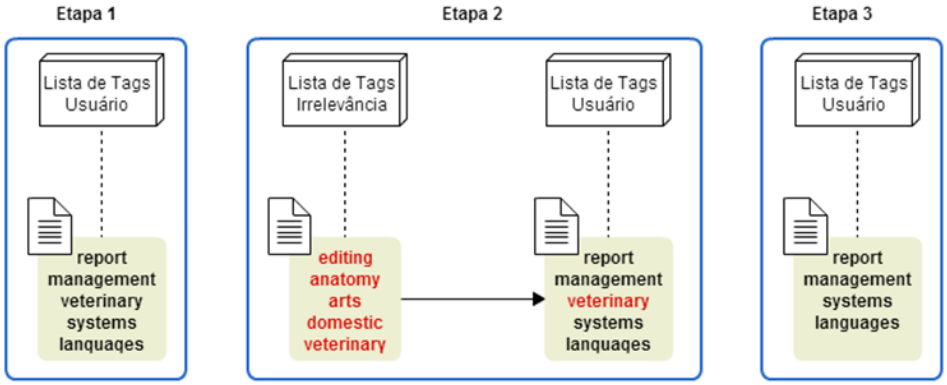
Esse tipo de problema pode ser agravado em um ambiente como o das bibliotecas. Alguns usuários efetuam empréstimos de livros para repassá-los para outras pessoas. Isso pode diminuir a acurácia do sistema de recomendação, pois o usuário não tem interesse no livro emprestado, o que gera uma sugestão de conteúdo incompatível com a necessidade do usuário.

Para amenizar esse tipo de problema, uma lista de tags irrelevantes é implementada para identificar termos que não são pertinentes. Inicialmente, o usuário recebe uma lista de livros recomendados com base em suas tags e avalia cada recomendação com uma nota de 0 a 3. Uma avaliação próxima de 3 indica maior interesse do usuário pelo item recomendado. Quando um livro é classificado pelo usuário com uma nota 0, então este item não é de interesse do usuário. Assim, uma recomendação irrelevante (Falso Positivo) é identificada por meio do feedback do usuário para formar a lista de tags irrelevantes. Cada usuário terá uma lista de tags irrelevantes, que será formada pelas tags dos livros que forem classificados com nota 0 na lista de recomendação.

Em seguida, quando novas recomendações forem feitas, a lista de tags do usuário é alterada com base na lista de irrelevância. As tags classificadas como irrelevantes são removidas da lista de tags do usuário para evitar recomendações com base em termos que não representam conteúdo de interesse do usuário. Como exemplo, observando a Fiura 3.2, na etapa 1 a lista de tags do usuário contém termos extraídos dos metadados de livros que foram emprestados. Quando o usuário

avalia um item com nota 0 por meio do seu feedback, forma-se na etapa 2, a “Lista de Tags Irrelevância” que contem as tags, editing, anatomy, arts, domestic e veterinary. A tag “veterinary” classificada como irrelevante é detectada e removida da “Lista de tags do usuário”. A partir desse momento, a “Lista de Tags Usuário” (etapa 3) não contém mais a tags responsáveis por gerar recomendações que não sejam pertinentes ao usuário. O fluxo de remoção da tag é representado na Figura 3.5.

Figura 3.5 – Fluxo de remoção da tag irrelevante



3.6.5 Lista de Tags de Área

Os sistemas de recomendação de conteúdo apresentam limitações que podem prejudicar a sua eficácia. Quando novos usuários acessam o sistema, as avaliações de itens geralmente não são efetuadas inicialmente, o que dificulta o processo de recomendação de conteúdo e a mensuração do grau de semelhança entre os usuários. Isso pode ser um problema recorrente em bibliotecas, pois alguns usuários não costumam efetuar empréstimos de livros durante algum tempo, o que dificulta a identificação de seus interesses.

As transferências de curso também impactam em mudanças de interesses, já que as disciplinas e temas abordados são distintos de acordo com cada período

do curso. Mesmo que o usuário não mude de curso ao longo dos períodos, as disciplinas possuem um conteúdo diferenciado, o que pode alterar o interesse e importância sobre os assuntos.

Para amenizar esse tipo de problema, utiliza-se uma lista de tags de área, que é formada pela extração das tags dos livros emprestados por usuários no período do curso. Assim, o perfil do curso em determinado momento no tempo pode ser identificado. Cada curso terá uma lista de tags dinâmicas por período que poderá sofrer alterações de acordo com os empréstimos efetuados por usuários que estejam cursando aquele período. Quando um usuário ingressa em um curso as tags referentes aos empréstimos do período do curso são direcionadas para a lista de tags do usuário, permitindo que o conteúdo abordado no período seja sugerido aos novos usuários.

Como exemplo, suponha que Alice, João e Carlos estão matriculados no terceiro período do curso de Administração e realizam vários empréstimos de livros. Com base nesses empréstimos a lista de tags desse curso no referido período é criada, e servirá para indicar conteúdo para o Marcos que irá cursar o terceiro período de Administração. Isso faz com que Marcos receba recomendações sobre conteúdo que geralmente é utilizado no período em que ele está matriculado. Esse processo pode auxiliar nas escolhas de Marcos, pois nos períodos anteriores ele não havia efetuado empréstimo, não tendo em sua lista nenhuma tag que pudesse representar suas preferências. Mas com base nas tags de área, foi possível gerar recomendações sobre temas que serão abordados no período e que provavelmente possam ser de seu interesse.

Assim como é feito na lista de tags do usuário, as tags da lista de área também recebem um peso que é calculado pelo método estatístico TF-IDF para identificar a relevância da tag na lista.

3.7 Agrupamento de Tags

Nessa etapa as listas de tags são comparadas para gerar as recomendações. Considera-se três grupos de corte definidos com base no peso das tags:

- Grupo 1: $\geq 70\%$;
- Grupo 2: $\geq 40\%$ e $\leq 69\%$;
- Grupo 3: $\leq 39\%$.

A tag pertence ao grupo 1 quando seu peso é 70% maior que as demais. Para que uma recomendação seja feita com base nesse grupo, no mínimo quatro tags extraídas de um livro ou documento devem ser idênticas as tags do grupo 1. Se essa condição não for satisfeita, analisa-se a tag extraída com base no grupo 2, neste caso no mínimo 5 tags devem ser compatíveis. Também, no grupo 3 no mínimo 5 tags que forem comparadas devem ser idênticas para que haja alguma recomendação.

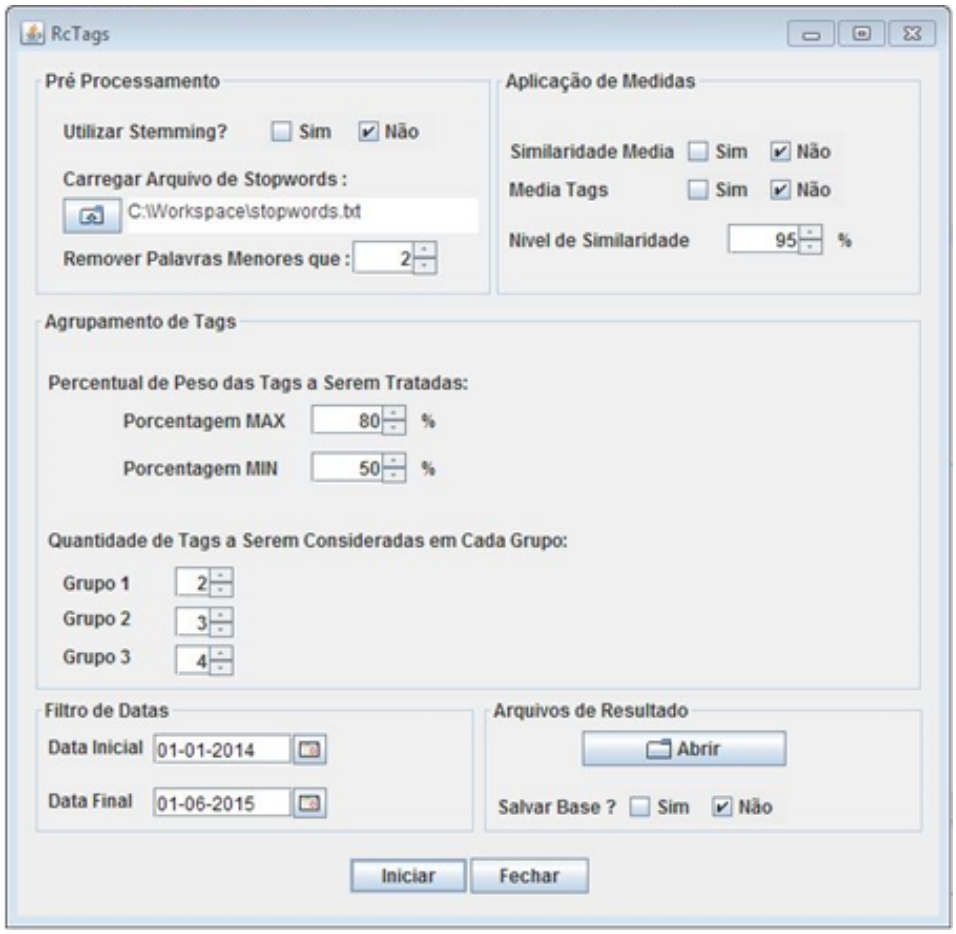
A quantidade de tags e percentuais dos grupos foram definidos observando os valores de precisão e revocação que obtiveram melhores resultados nos experimentos. Para isso, o algoritmo foi executado várias vezes variando os parâmetros de percentuais dos grupos e de quantidade de tags para identificar os melhores valores de corte.

Na Figura 3.6 é apresentada a tela do sistema que foi desenvolvido para facilitar a variação dos parâmetros a cada execução do algoritmo.

3.8 Obtendo o Perfil do Usuário

Para traçar o perfil do usuário são utilizadas duas estratégias: mapeamento do seu histórico de empréstimos e coleta de feedback. Primeiramente, os metadados dos livros que o usuário emprestou são extraídos para formar uma lista de tags

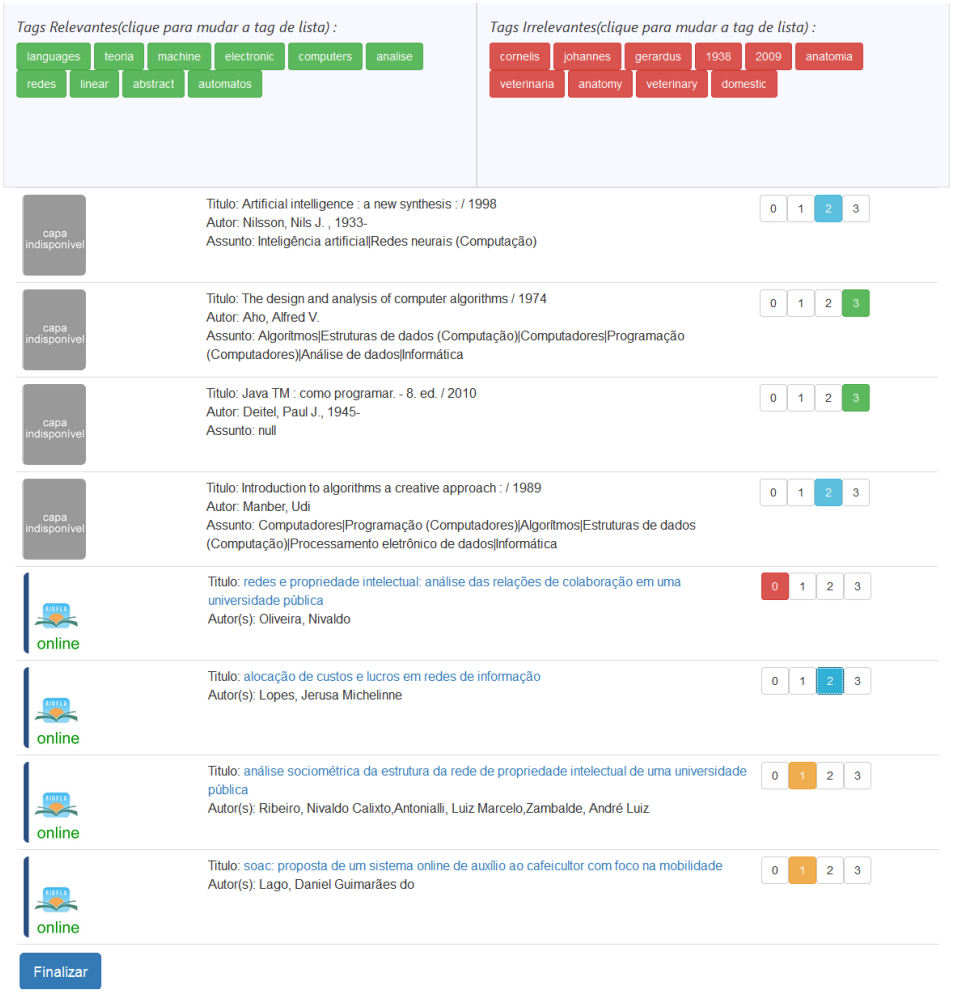
Figura 3.6 – Tela de parâmetros para execução do algoritmo



que irá representar seus interesses. Assim, cada usuário possui uma lista de tags que é utilizada para gerar recomendações de livros e documentos.

Na segunda etapa, uma lista de itens formados por livros e documentos é submetida aos usuários, que envia seu feedback ao avaliar cada item de acordo com seu interesse. A lista de tags do usuário é alterada de acordo com as avaliações. Assim, ao combinar informações do seu registro de empréstimos e dados coletados pelo feedback, torna-se possível obter um perfil completo e confiável de acordo interesse do usuário.

Figura 3.7 – Tela com lista de itens recomendados que o usuário deve avaliar (feedback)



A Figura 3.7 representa a tela com as recomendações que o usuário deve avaliar. Cada usuário possui sua lista de recomendações personalizadas. O acesso a essas informações é feito por meio de um link enviado para o e-mail de cada usuário que recebeu recomendações. A mensagem possui o link para acesso a sua lista de recomendações e informações que devem ser consideradas para classificar os itens. Nesse caso, o usuário deve avaliar cada item com uma nota de 0 a 3, sendo:

- Valor 0, significa que o item recomendado não interessa para o usuário;
- Valor 1, item recomendado interessa pouco para o usuário;
- Valor 2, item recomendado interessa ao usuário;
- Valor 3, usuário tem muito interesse no item recomendado.

Caso o item seja avaliado com uma nota 0, considera-se que a recomendação não deveria ter sido sugerida para o usuário. Assim, os metadados extraídos desse item são transformados em tags para ser enviada à lista de tags irrelevantes do usuário. Na próxima recomendação os itens que tiverem alguma relação com essas tags consideradas irrelevantes, não serão mais indicados para o usuário.

No topo da página o usuário tem acesso a sua lista de tags relevantes e irrelevantes, e pode realocá-las entre as listas de acordo com seu interesse. Caso o documento recomendado seja proveniente do repositório, ao clicar sobre o seu título, o usuário é redirecionado para a página que contem informações completas que descrevem detalhadamente o conteúdo do item a ser classificado.

4 EXPERIMENTOS E AVALIAÇÕES

Nos tópicos seguintes, um exemplo de execução do algoritmo é descrito, passando por todas as fases já descritas anteriormente. Em seguida, os resultados são avaliados com base nas métricas mencionadas nos capítulos anteriores.

4.1 Formação da Base de Experimento

Para realizar essa pesquisa foram coletadas informações da base de dados do sistema de gerenciamento da biblioteca da universidade e da base de dados do repositório institucional. A base da biblioteca contém os registros das atividades realizadas na biblioteca, como o histórico dos empréstimos, informações sobre usuários e livros. Já os dados provenientes do repositório institucional são referentes a documentos digitais, que contém informações de artigos, dissertações e teses.

Para a base de dados da biblioteca, foram considerados 27.767 empréstimos do período de 01/01/2014 a 01/06/2015, em que 1.769 usuários receberem recomendações e 122 pessoas enviaram seu feedback. Muitos usuários utilizam o e-mail pessoal, por isso não acessam o link da lista de recomendações que é enviada para o e-mail institucional. No relatório de Auto avaliação Institucional desenvolvido pela Comissão Própria de Avaliação da UFLA – CPA constatou-se que somente 54,7% dos alunos utilizam o e-mail institucional (SCOLFORO et al., 2016). A falta de interesse dos usuários em participar da pesquisa e classificar as recomendações também prejudica a coleta de dados.

Tabela 4.1 – Informações da base de dados da biblioteca

Base de Dados Bibliotecas			**Base de Dados Repositório**		
Livros	Tags	Tags (filtro)	Documentos	Tags	Tags (filtro)
1,795	988,848	363,392	9,478	693,089	495,259

Conforme tabela 4.1, após a fase de processamento a quantidade de tags recuperadas da base da biblioteca diminui para 363.392 e as tags recuperadas do repositório após o tratamento reduz para 495.259, sem considerar termos repetidos.

Os usuários envolvidos na pesquisa fazem parte de 6 cursos distintos: mestrado em Ciência da Computação, graduação em Ciência da Computação, Sistemas de Informação, Administração, Zootecnia e Veterinária. Propositalmente, para verificar a consistência e avaliar as recomendações do algoritmo com maior critério, foram escolhidos cursos que possuem conteúdos diferenciados.

As listas de recomendações contem 30 itens, entre livros e documentos que cada usuário deve avaliar. Um maior número de itens pode tornar o processo de avaliação dispendioso, causando desistências no momento da classificação.

4.2 Execução do Experimento

Para iniciar os experimentos, informações da base de dados da biblioteca e do repositório são integradas e coletadas para construção do perfil do usuário. O histórico de empréstimos do usuário na biblioteca é utilizado para gerar seu perfil a partir da extração dos metadados dos livros. Os documentos do repositório, assim como os livros, são recomendados para avaliação na fase de feedback.

Na etapa de pré-processamento as informações coletadas são submetidas a vários tratamentos de filtragem para formar as listas de tags que são utilizadas para gerar as recomendações. Aplica-se a técnica do TF-IDF que estabelece pesos às tags para identificar o grau de importância do termo para o usuário. As listas de tags de livros e a lista de tags de documentos são formadas sem pesos, pois não é necessário identificar a importância do termo no documento, mas qual sua importância para o usuário. Para exemplificar, a tabela 4.2 possui o conjunto de livros que foram emprestados por um usuário denominado UsuárioAluno.

Tabela 4.2 – Livros emprestados por um usuário

Usuário	Livros emprestados
UsuárioAluno	APLICACOES DE COMPUTADORES INTRODUCAO A PROGRAMACAO LINEAR
	ORGANIZACAO ESTRUTURADA DE COMPUTADORES STRUCTURED COMPUTER ORGANIZATION
	ELEMENTOS DE TEORIA DA COMPUTACAO ELEMENTS OF THE THEORY OF COMPUTATION
	LINGUAGENS FORMAIS E AUTOMATOS LIVROS DIDATICOS INFORMATICA UFRGS BOOKMAN
	PESQUISA OPERACIONAL
	FUNDAMENTOS MATEMATICOS PARA A CIENCIA DA COMPUTACAO
	LINEAR PROGRAMMING AND NETWORK FLOWS
	INTRODUCTION TO ALGORITHMS
	TEXTBOOK OF VETERINARY ANATOMY FOURTH EDITION TRATADO DE ANATOMIA VETERINARIA
	ALGORITMOS E SEUS FUNDAMENTOS

Na tabela 4.3 estão as tags que foram extraídas dos metadados desses livros sendo identificados os seus pesos de importância definidos a partir da aplicação do calculo do TF-IDF.

Na tabela 4.4 é apresentada as listas de tags da área que contém os termos dos livros emprestados em determinado período de um curso específico, que neste exemplo é ciência da computação. O objetivo dessa lista é identificar as tags do curso que o usuário está matriculado de acordo com os empréstimos feitos a cada período.

Com as listas de tags definidas, a próxima etapa trata do agrupamento dessas informações para gerar as recomendações. Assim, as tags são separadas em três grupos de acordo com seu peso: o primeiro grupo é formado pelas tags que tiverem valor maior ou igual a 80% do maior peso, o segundo grupo pelas tags que tiverem peso entre 79% e 50%, e o terceiro grupo pelas tags com peso menor ou igual a 49%. Com os grupos formados considera-se os seguintes critérios para gerar as recomendações:

Tabela 4.3 – Tags extraídas dos livros emprestados pelo usuário

Usuário	Tags do usuário	Peso
UsuárioAluno	logic	0.0662
	languages	0.0191
	electronic	0.0514
	computers	0.0410
	teoria	0.0164
	logica	0.0637
	computadores	0.0398
	machine	0.0140
	data	0.0498
	matematica	0.0634
	processing	0.0494
	linear	0.0846
	models	0.0524
	algebra	0.0613
	complexity	0.0098
	computational	0.0098

Tabela 4.4 – Tags referentes ao curso e seus períodos

Curso	Período	Tag - peso
Mestrado em Ciência da Computação	Segundo Período	architecture - 0.1397
		computer - 0.1397
		arquitetura - 0.1397
		computador - 0.1397
		computadores - 0.1397
Mestrado em Ciência da Computação	Terceiro Período	linear - 0.1590
		programming - 0.1590
		programacao - 0.1590
		algebra - 0.1128
		logica - 0.1128
Mestrado em Ciência da Computação	Quarto Período	analise - 0.1505
		numerica - 0.1505
		numerical - 0.1505
		analysis - 0.1505
		estatistica - 0.1128

- Se no mínimo duas tags do grupo 1 da lista do usuário forem iguais as tags do livro, então o livro/documento é recomendado para o usuário.

- Se não existir tags iguais no grupo 1, verifica-se no grupo 2 se existem no mínimo 3 tags iguais a do livro para que seja recomendado.

- Se no conjunto de tags do livro/documento não existir nenhuma tag no grupo 1 nem no grupo 2, então verifica se existe no grupo 3. Se no mínimo quatro tags do livro correspondam com as tags do usuário, então o livro/documento é recomendado.

Esse esquema de discretização foi adotado para evitar recomendações que possam ser irrelevantes para o usuário, pois uma tag que possui um menor valor de peso não acrescenta tanta importância quanto uma tag que possua um alto valor de peso. Assim, se uma recomendação é feita considerando apenas uma tag de menor valor, a probabilidade de irrelevância da recomendação seria maior. Por isso, as tags que possuem menor peso precisam ser combinadas para agregar mais valor em uma recomendação.

Por meio do calculo de similaridade identifica-se o grau de semelhança entre os usuários. Assim, livros que foram emprestados por eles podem ser indicados para outros usuários que possuam um grau alto grau de semelhança. Para identificar a proximidade entre dois usuários com base em sua lista de tags, considera-se a seguinte situação hipotética.

O usuário A possui em sua lista a tag X com peso 0.3, a tag Y com peso 0.0 e a tag Z com peso 0.5, enquanto que o usuário B possui a tag X com peso 0.5, a tag Y com peso 0.4 e a tag Z com peso 0.3. Cada uma das tags faz parte da lista de tags do usuário com o peso calculado pelo TF-IDF. Aplicando os valores na fórmula da similaridade em (4.1) tem-se o seguinte:

$$similaridade(U_A, U_B) = \frac{(0.3 \times 0.5) + (0.0 \times 0.4) + (0.5 \times 0.3)}{\sqrt{0.3^2 + 0.0^2 + 0.5^2} \times \sqrt{0.5^2 + 0.4^2 + 0.3^2}} = 0.73 \tag{4.1}$$

O resultado do calculo de similaridade entre o usuário A e o usuário B, resultou em um grau de semelhança entre eles de 0.73. Com todos os resultados calculados a matriz $U_A \times U_B$ é criada como na tabela 4.5.

Tabela 4.5 – Matriz de usuários $U_A \times U_B$

	Usuário A	**Usuário B**
Usuário A	1.0	0.73
Usuário B	0.73	1.0

Com base nesse cálculo, os usuários que possuam alto valor de similaridade (95%) recebem indicações de livros que foram emprestados por eles. Com isso, uma lista de livros e documentos é gerada como recomendação para o usuário avaliar com uma nota de 0 (recomendação não interessa) a 3 (recomendação interessa). Com base no feedback, a lista de tags irrelevantes é formada com as tags dos livros/documentos recomendados, mas que o usuário avaliou com nota 0. A lista de tags do usuário é comparada com a lista de tags irrelevantes, se alguma tag irrelevante aparecer em sua lista então será removida para evitar recomendações de conteúdo irrelevantes. Essas etapas vão sendo repetidas em um ciclo constante, quanto mais empréstimos e avaliações o usuário fizer, maior será a acurácia das recomendações geradas pelo sistema.

4.3 Avaliação do Experimento

Para avaliar as recomendações utiliza-se as medidas de precisão e revocação apresentadas na seção anterior. A avaliação do experimento consiste em duas etapas: na primeira etapa são avaliadas as recomendações com base no histórico de empréstimos do usuário, na segunda etapa utiliza-se o feedback do usuário para avaliar as recomendações que foram sugeridas. Considera-se como relevantes os livros emprestados pelo usuário e que o sistema recomendou em sua lista. Irrelevantes são livros recomendados para o usuário, mas não foram emprestados por

ele. Sendo assim, a matriz de confusão é preenchida para realizar os cálculos de precisão e recall, considerando que:

- N_{rr} = livros recomendados que foram emprestados pelo usuário.
- N_{ir} = livros recomendados, mas não emprestados pelo usuário.
- N_{rn} = livros não recomendados, mas emprestados pelo usuário.
- N_{in} = não se aplica, pois não há como identificar livros não recomendados e que também não foram emprestados pelo usuário, já que a lista de recomendação é gerada com base nos empréstimos do usuário em um determinado período de tempo. Esse valor é utilizado se considerada toda a base de dados da biblioteca, porém o aumento do tempo de processamento tornaria inviável a realização dos experimentos nessa pesquisa.

A execução do experimento 5, que considerou 4, 5 e 5 tags, obteve melhores resultados, com o valor de precisão 0.4490, revocação 0.4352 e F-Score 0.4420, conforme apresentado na tabela 4.6:

Inicialmente para a medida de corte utilizada considerou-se um valor exato de tags. Porém, muitos dados eram desconsiderados ou recuperados além do necessário. Por isso, aplicou-se a separação percentual dos grupos com base nos pesos das tags, e os limites de corte foram definidos com base em Durao e Dolog (2009), que utiliza valores próximos de 70%. Um valor muito acima deste limite pode omitir recomendações interessantes para o usuário, enquanto que valores muito abaixo podem gerar muitas recomendações desnecessárias.

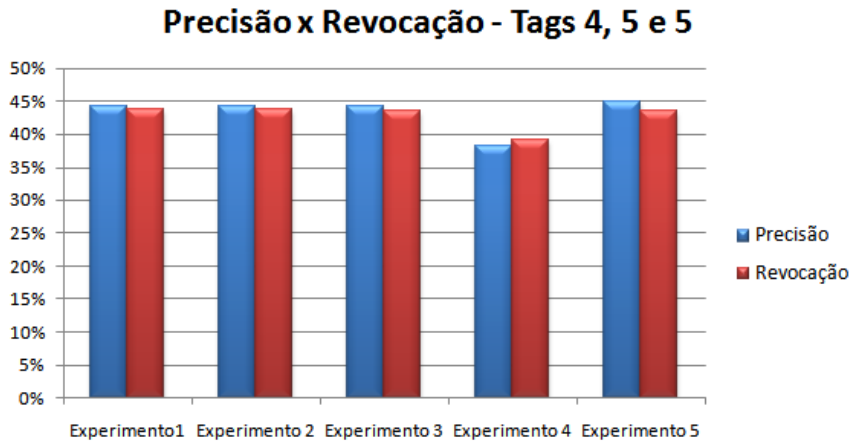
Na Figura 4.1 estão representados os experimentos das execuções do algoritmo considerando 4, 5 e 5 tags, sendo possível analisar o melhor ponto de corte das execuções que apresentaram melhores resultados, neste caso obtido pelo experimento 5.

Analisando os grupos de tags detalhadamente no experimento 5 na Figura 4.2, verifica-se que o grupo 4, 5 e 5 de tags obteve melhores resultados do

Tabela 4.6 – Resultado das execuções algoritmo variando os parâmetros para cada experimento

Experimentos	**Grupos**				
Experimento 1	Grupo 1: $\geq 90\%$ Grupo 2: $\geq 60\%$ e $\leq 89\%$ Grupo 3: $\leq 59\%$				
Tags	1, 2 e 3	2, 3 e 4	3, 4 e 4	4, 5 e 5	2, 4 e 6
Precisão	0.3421	0.3363	0.4081	0.4412	0.4271
Revocação	0.4375	0.3908	0.4375	0.4372	0.4375
F-Score	0.3839	0.3615	0.4223	0.4392	0.4322
Experimento 2	Grupo 1: $\geq 85\%$ Grupo 2: $\geq 55\%$ e $\leq 84\%$ Grupo 3: $\leq 54\%$				
Tags	1, 2 e 3	2, 3 e 4	3, 4 e 4	4, 5 e 5	2, 4 e 6
Precisão	0.3401	0.3362	0.4094	0.4424	0.4244
Revocação	0.4375	0.3908	0.4375	0.4372	0.4375
F-Score	0.3827	0.3615	0.4230	0.4398	0.4308
Experimento 3	Grupo 1: $\geq 80\%$ Grupo 2: $\geq 60\%$ e $\leq 79\%$ Grupo 3: $\leq 59\%$				
Tags	1, 2 e 3	2, 3 e 4	3, 4 e 4	4, 5 e 5	2, 4 e 6
Precisão	0.3416	0.3340	0.4077	0.4412	0.4228
Revocação	0.4355	0.3908	0.4355	0.4352	0.4355
F-Score	0.3829	0.3602	0.4211	0.4382	0.4291
Experimento 4	Grupo 1: $\geq 80\%$ Grupo 2: $\geq 50\%$ e $\leq 79\%$ Grupo 3: $\leq 49\%$				
Tags	1, 2 e 3	2, 3 e 4	3, 4 e 4	4, 5 e 5	2, 4 e 6
Precisão	0.3469	0.3921	0.3541	0.3825	0.3658
Revocação	0.4270	0.4270	0.3908	0.3906	0.3908
F-Score	0.3828	0.4088	0.3715	0.3865	0.3779
Experimento 5	Grupo 1: $\geq 70\%$ Grupo 2: $\geq 40\%$ e $\leq 69\%$ Grupo 3: $\leq 39\%$				
Tags	1, 2 e 3	2, 3 e 4	3, 4 e 4	4, 5 e 5	2, 4 e 6
Precisão	0.2889	0.3336	0.3564	0.4490	0.4120
Revocação	0.3908	0.3908	0.3908	0.4352	0.4355
F-Score	0.3322	0.3600	0.3729	0.4420	0.4234

Figura 4.1 – Gráfico com os valores de precisão e revocação para cada experimento considerando o grupo com 4, 5 e 5 tags

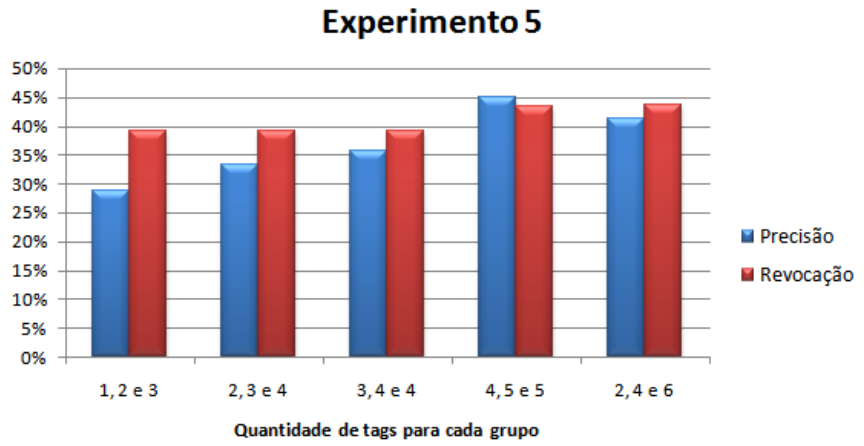


que os demais, pois algumas tags quando analisadas conjuntamente agregam um significado mais contundente ao conteúdo. Um termo composto analisado separadamente perde o significado da representação do conteúdo. Por exemplo, o termo "arquitetura de computadores" forma três tags, o termo "de" é removido no pré processamento, e os termos "arquitetura" e "computadores" são analisados separadamente. Se apenas um termo pertence a lista de tags há perda de significado, pois os termos "arquitetura" e "computadores" quando analisados separadamente possuem conceitos variados.

Para melhorar a precisão das recomendações, a quantidade de tags usadas para comparação deve aumentar até certo limite, como foi feito no experimento 5 no grupo 4, 5 e 6 da Figura 4.2. Ao aumentar o número de tags para comparação, a precisão tende a aumentar e o valor de revocação pode diminuir. Mesmo assim, o valor de revocação manteve um valor alto em relação aos demais grupos, sendo a quantidade de tags 4, 5 e 5 aplicada para gerar as recomendações.

Ao encontrar os melhores limites de corte o algoritmo é executado novamente para gerar a lista de recomendações (feedback) que será enviada para o

Figura 4.2 – Gráfico com os percentuais de precisão e revocação analisados para todos os cinco grupos de tags no experimento 5

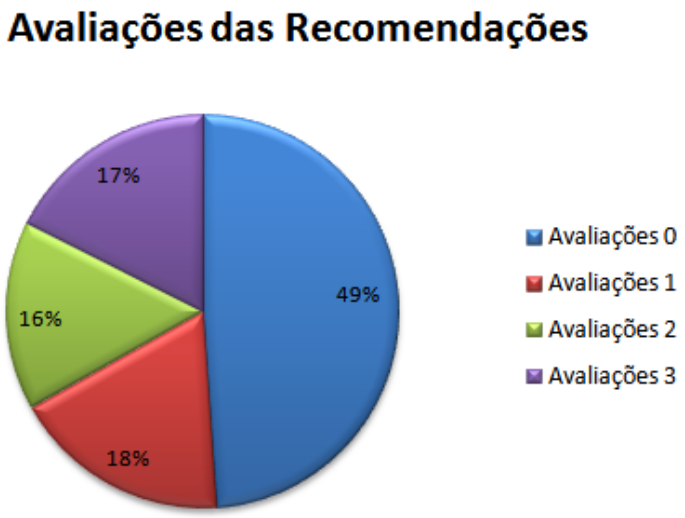


usuário. Nesta segunda etapa, a satisfação do usuário é medida em relação às recomendações feitas por meio do seu feedback. Foram coletadas 3.666 avaliações.

Pela Figura 4.3, 49% das avaliações feitas pelos usuários indicam que as sugestões de conteúdo não são de seu interesse. No entanto, 51% do conteúdo recomendado foi avaliado com notas 1, 2 ou 3, o que indica que esses itens sugeridos foram bem aceitos pelo usuário. Nessa primeira etapa apenas o histórico de empréstimos do usuário foi utilizado para gerar as recomendações. Os valores inicialmente encontrados não são ideais para um sistema de recomendação eficiente, mas são propícios para iniciar as primeiras sugestões de conteúdo, e podem ser melhorados na próxima etapa conforme as avaliações dos usuários vão sendo coletadas.

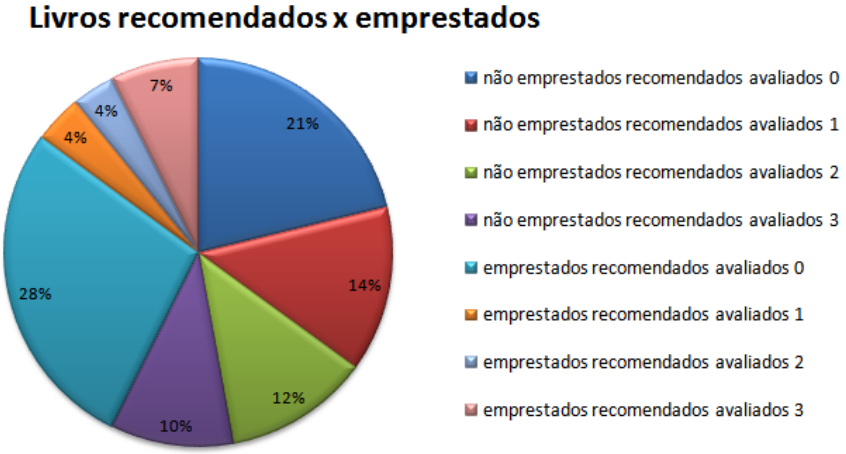
Ao analisar os dados da primeira etapa de recomendação e relacioná-los com os dados de feedback do usuário, verifica-se pela Figura 4.4 que 28% dos livros que o usuário emprestou foram recomendados mas tiveram avaliações negativas. Isso ocorre, pois alguns livros que o usuário faz empréstimos nem sempre são de seu interesse, ou suas preferências variam com o passar do tempo. Dos livros emprestados recomendados, 15% (4%, 4% e 7%) são classificados pelos

Figura 4.3 – Gráfico com os percentuais das avaliações feitas pelos usuários sobre cada livro/documento recomendado



usuários como interessantes, o que demonstra que o comportamento avaliativo dos usuários nesses itens está de acordo com suas opções de empréstimo e foram sugeridos corretamente.

Figura 4.4 – Gráfico com os percentuais de aceitação das recomendações de livros que foram ou não emprestados pelos usuários



Observa-se que 36% (14%, 12% e 10%) dos livros que ainda não foram emprestados tiveram uma boa aceitação pelos usuários. Isso demonstra que o al-

goritmo conseguiu sugerir uma parcela significativa de livros que eram desconhecidos pelos usuários, mas que foram avaliados de forma positiva, demonstrando interesse pelas recomendações.

Com base na tabela 4.7, calcula-se os valores para precisão, revocação e f-score, que são respectivamente 34.9%, 29% e 31%. Verifica-se que o valor de f-score adquirido pelo feedback do usuário ficou próximo do valor de f-score (44%) calculado na primeira etapa ao utilizar o histórico de empréstimos. Assim, temos que as recomendações baseadas no histórico de fato representam o interesse do usuário.

Tabela 4.7 – Valores de feedback dos usuários

	Emprestado	**Não emprestado**
Relevante	547	1326
Irrelevante	1016	777

5 CONCLUSÃO E TRABALHOS FUTUROS

Em um ambiente acadêmico a disponibilização e diversificação de informações é fundamental para garantir a qualidade e desenvolvimento das pesquisas. Alguns sistemas são utilizados para este fim, porém nem sempre proporcionam tecnologias inteligentes para facilitar a busca por conteúdo de forma integrada, como é o caso dos repositórios digitais e os sistemas de gerenciamento de bibliotecas. Por isso, sistemas de recomendação de conteúdo são aplicados em ambientes acadêmicos para auxiliar os usuários na busca por informações novas e relevantes. A integração desses serviços e utilização de técnicas de recomendação de conteúdo sobre essas tecnologias geram benefícios que permitem a disponibilização de recursos relevantes para os usuários e a disseminação direcionada de diversos tipos de conteúdo.

Nesse sentido, é proposto um sistema de recomendação que aplica uma abordagem híbrida e utiliza listas de tags para modelar o perfil do usuário, adaptando-se às mudanças de interesse ao longo do tempo que permita a indicação de conteúdo novo e relevante. O sistema extrai da base de dados da biblioteca os metadados dos livros emprestados e os transforma em tags que são utilizadas para modelar o perfil do usuário, que recebe sugestões de conteúdo para ser avaliada de acordo com seu interesse. Os itens da lista de recomendação são constituídos por livros da base da biblioteca e documentos que são recuperados do repositório institucional. Essa integração proporcionou mais conteúdo ao usuário, mas de forma direcionada e controlada de acordo com seu perfil. Por meio do feedback do usuário foi possível melhorar a precisão das próximas recomendações. Além disso, a utilização das listas de tags dinâmicas possibilitou a identificação das mudanças de comportamento dos usuários ao longo do tempo.

O sistema proporcionou sugestões de itens que anteriormente não eram explorados, mas foram classificados pelos usuários como recomendações importantes. Identificou-se que alguns usuários que efetuaram empréstimos de livros

perderam o interesse pelo mesmo conteúdo. Porém, por meio do feedback e das listas de tags da área, a lista de tags do usuário foi ajustada de acordo com as variações de interesse ao longo do tempo, gerando recomendações mais precisas, adaptadas às novas preferências do usuário.

Novas abordagens podem ser exploradas nesse estudo para melhorar a precisão das recomendações, como a utilização de ontologias e semântica, que melhoram a qualidade da representação dos termos nas listas de tags. Além disso, a aplicação proposta pode ser acoplada em um sistema de busca integrado, que mapeia as interações em tempo real e recupera os dados de pesquisa de usuários utilizando-os para agregar conhecimento para pesquisas futuras. No que se refere a avaliação das recomendações, outras medidas podem ser utilizadas sobre outros aspectos, como erro médio absoluto (MAE) e curva ROC, para determinar melhores classificações e mensurar a performance do sistema de recomendação.

REFERÊNCIAS

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. **IEEE Transactions on Knowledge and Data Engineering**, v. 17, n. 6, p. 734–749, June 2005. ISSN 1041-4347.

ALVARES, R. V. **ALGORITMOS DE STEMMING E O ESTUDO DE PROTEOMAS**. Tese (Tesde de Doutorado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

BARRAGáNS-MARTÍNEZ, A. B. et al. A hybrid content-based and item-based collaborative filtering approach to recommend tv programs enhanced with singular value decomposition. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 180, n. 22, p. 4290–4311, nov. 2010. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2010.07.024>.

CASTRO-HERRERA, C. A hybrid recommender system for finding relevant users in open source forums. In: **Managing Requirements Knowledge (MARK), 2010 Third International Workshop on**. [S.l.: s.n.], 2010. p. 41–50.

CHIRITA, P. A. et al. P-tag: Large scale automatic generation of personalized annotation tags for the web. In: **Proceedings of the 16th International Conference on World Wide Web**. New York, NY, USA: ACM, 2007. (WWW '07), p. 845–854. ISBN 978-1-59593-654-7. Disponível em: <http://doi.acm.org/10.1145/1242572.1242686>.

DOLOG, P. et al. Recommending open linked data in creativity sessions using web portals with collaborative real time environment. v. 17, n. 12, p. 1690–1709, aug 2011. <|>http://www.jucs.org/jucs$_1$7$_1$2/$recommending_o pen_l inked_d ata$|.

DSPACE. **DSPace 4.x Documentation**. 2014. Disponível em: <https://wiki.duraspace.org/display/DSPACE/Home>. Acesso em: 10 jan. 2016.

DURAO, F.; DOLOG, P. Social and behavioral aspects of a tag-based recommender system. In: **2009 Ninth International Conference on Intelligent Systems Design and Applications**. [S.l.: s.n.], 2009. p. 294–299. ISSN 2164-7143.

DURAO, F.; DOLOG, P. Extending a hybrid tag-based recommender system with personalization. In: **Proceedings of the 2010 ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2010. (SAC '10), p. 1723–1727. ISBN 978-1-60558-639-7. Disponível em: <http://doi.acm.org/10.1145/1774088.1774457>.

DURAO, F.; DOLOG, P. Improving tag-based recommendation with the collaborative value of wiki pages for knowledge sharing. **Journal of Ambient**

Intelligence and Humanized Computing, v. 5, n. 1, p. 21–38, 2014. ISSN 1868-5145. Disponível em: <http://dx.doi.org/10.1007/s12652-012-0119-x>.

DURAO, F. A.; DOLOG, P. A personalized tag-based recommendation in social web systems. **CoRR**, abs/1203.0332, 2012. Disponível em: <http://arxiv.org/abs/1203.0332>.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados**. sixth. [S.l.]: Pearson, 2011. ISBN 978-85-7936-085-5.

GAMA, J. **Knowledge discovery from data streams**. Boca Raton, FL: CRC Press, 2010. (Chapman & Hall/CRC data mining and knowledge discovery series). ISBN 978-1-439-82611-9. Disponível em: <http://opac.inria.fr/record=b1130806>.

GARCIA, C. A.; FROZZA, R. Sistema de recomendação de produtos utilizando mineração de dados. **Tecno-Lógica**, v. 17, n. 1, p. 78–90, feb 2013.

JI, K.; SHEN, H. Addressing cold-start: Scalable recommendation with tags and keywords. **Knowledge-Based Systems**, v. 83, p. 42 – 50, 2015. ISSN 0950-7051. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705115001008>.

JIANG, Y. G. et al. Representations of keypoint-based semantic concept detection: A comprehensive study. **IEEE Transactions on Multimedia**, v. 12, n. 1, p. 42–53, Jan 2010. ISSN 1520-9210.

JR, C. V. B.; OLIVEIRA, M. A. de. Recommender systems in social networks. **JISTEM - Journal of Information Systems and Technology Management**, scielo, v. 8, p. 681 – 716, 12 2011. ISSN 1807-1775. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-17752011000300009&nrm=iso>.

JUNG, C. F. **Metodologia para pesquisa e desenvolvimento: aplicada a novas tecnologias, produtos e processos**. [S.l.]: Axcel Books, 2004.

KARDAN, A. A.; EBRAHIMI, M. A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 219, p. 93–110, jan. 2013. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2012.07.011>.

KIM, H.-N. et al. Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. **Electronic Commerce Research and Applications**, v. 9, n. 1, p. 73 – 83, 2010. ISSN 1567-4223. Special Issue: Social Networks and Web 2.0. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1567422309000544>.

LAI, C.-H.; LIU, D.-R.; LIN, C.-S. Novel personal and group-based trust models in collaborative filtering for document recommendation. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 239, p. 31–49, ago. 2013. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2013.03.030>.

LIN, W.-C. et al. Keypoint selection for efficient bag-of-words feature generation and effective image classification. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 329, n. C, p. 33–51, fev. 2016. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2015.08.021>.

LIU, D.-R. et al. Recommending blog articles based on popular event trend analysis. **Information Sciences**, v. 305, p. 302 – 319, 2015. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S002002551500081X>.

LUO, H. L.; WEI, H.; LAI, L. L. Creating efficient visual codebook ensembles for object categorization. **IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans**, v. 41, n. 2, p. 238–253, March 2011. ISSN 1083-4427.

MARLOW, C. et al. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: **Proceedings of the Seventeenth Conference on Hypertext and Hypermedia**. New York, NY, USA: ACM, 2006. (HYPERTEXT '06), p. 31–40. ISBN 1-59593-417-0. Disponível em: <http://doi.acm.org/10.1145/1149941.1149949>.

MEDEIROS, S. A. **Gestão do conhecimento na sociedade da informação: repositório institucional da Universidade Federal de Lavras**. Dissertação (Mestrado) — Universidade Federal de Lavras, Lavras, MG, Brasil, 2012.

MORENO, M. N. et al. Web mining based framework for solving usual problems in recommender systems. a case study for movies recommendation. **Neurocomputing**, v. 176, p. 72 – 80, 2016. ISSN 0925-2312. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World ProblemsSelected papers from the {HAIS} 2013 conference. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231215005457>.

OLIVEIRA, A. P. S. de; COELLO, J. M. A. Desenvolvimento de algoritmo híbrido para sistemas de recomendação: filtragem colaborativa e etiquetagem social. **Anais do III Encontro de Iniciação em Desenvolvimento Tecnológico e Inovação**, set 2013.

PERGAMUM. **Tecnologia e características gerais**. 2015. Disponível em: <http://www.pergamum.pucpr.br/>. Acesso em: 10 dez. 2015.

POLATIDIS, N.; GEORGIADIS, C. K. A multi-level collaborative filtering method that improves recommendations. **Expert Syst. Appl.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 48, n. C, p. 100–110, abr. 2016. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2015.11.023>.

REATEGUI, E. B.; CAZELLA, S. C.; OSóRIO, F. S. Personalização de páginas web através dos sistemas de recomendação. **Tópico em Sistemas Interativos e Colaborativos. São Carlos**, 2006.

SALTER, J.; ANTONOPOULOS, N. Cinemascreen recommender agent: Combining collaborative and content-based filtering. **IEEE Intelligent Systems**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 21, n. 1, p. 35–41, jan. 2006. ISSN 1541-1672. Disponível em: <http://dx.doi.org/10.1109/MIS.2006.4>.

SANCHEZ, J. L. et al. Choice of metrics used in collaborative filtering and their impact on recommender systems. In: **2008 2nd IEEE International Conference on Digital Ecosystems and Technologies**. [S.l.: s.n.], 2008. p. 432–436. ISSN 2150-4938.

SCOLFORO, J. R. S. et al. **Relatório de Autoavaliação Institucional Referente ao ano de 2015: Primeiro Relatório Parcial do Triênio 2015-2017**. Lavras, 2016.

SERRANO-GUERRERO, J. et al. A google wave-based fuzzy recommender system to disseminate information in university digital libraries 2.0. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 181, n. 9, p. 1503–1516, maio 2011. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2011.01.012>.

SMITH, M. et al. Dspace: An open source dynamic digital repository. **D-Lib Magazine**, v. 9, n. 1, 2003. Disponível em: <http://dblp.uni-trier.de/db/journals/dlib/dlib9.html#SmithBBMWBST03>.

TEJEDA-LORENTE, A. et al. A quality based recommender system to disseminate information in a university digital library. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 261, p. 52–69, mar. 2014. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/j.ins.2013.10.036>.

USHARANI, J.; IYAKUTTI, K. A genetic algorithm based on cosine similarity for relevant document retrieval. **International Journal of Engineering Research and Technology**, v. 2, fev 2013.

VAISHNAVI, V.; KUECHLER, W. Design Research in Information Systems. 2004.

WILLETT, P. The porter stemming algorithm: then and now. **Program**, v. 40, n. 3, p. 219–223, 2006. Disponível em: <http://dx.doi.org/10.1108/00330330610681295>.

XAVIER, B. M.; SILVA, A. D. da; GOMES, G. R. R. Análise comparativa de algoritmos de redução de radicais e sua importância para a mineração de texto. **Pesquisa Operacional para o Desenvolvimento**, v. 5, n. 1, p. 84–99, jan 2013.

ZHU, S. et al. Scaling up top-k cosine similarity search. **Data Knowledge Engineering**, v. 70, n. 1, p. 60 – 83, 2011. ISSN 0169-023X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169023X10001114>.