ULBRA CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS
COMUNIDADE EVANGÉLICA LUTERANA "SÃO PAULO"
Recredenciado pela Portaria Ministerial nº 3.607 - D.O.U. nº 202 de 20/10/2005

Bruno Bandeira Fernandes

DESENVOLVIMENTO DE UM MECANISMO DE RECOMENDAÇÃO DE PUBLICAÇÕES RELACIONADAS PARA A PLATAFORMA *WORDPRESS*

Palmas – TO

2016

Bruno Bandeira Fernandes

DESENVOLVIMENTO DE UM MECANISMO DE RECOMENDAÇÃO DE PUBLICAÇÕES RELACIONADAS PARA A PLATAFORMA *WORDPRESS*

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso II (TCC II) do curso de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. Dr. Edeilson Milhomem da Silva

Palmas – TO

2016

Bruno Bandeira Fernandes

DESENVOLVIMENTO DE UM MECANISMO DE RECOMENDAÇÃO DE *PUBLICAÇÕES* RELACIONADAS PARA A PLATAFORMA WORDPRESS

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso II (TCC II) do curso de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. Dr. Edeilson Milhomem da Silva

Aprovada em Junho de 2016.

BANCA EXAMINADORA

_____

Prof. Dr. Edeilson Milhomem da Silva

Centro Universitário Luterano de Palmas

_____

Prof. M.Sc. Jackson Gomes de Souza

Centro Universitário Luterano de Palmas

_____

Prof. M.Sc. Parcilene Fernandes de Brito

Centro Universitário Luterano de Palmas

Palmas – TO

2016

AGRADECIMENTOS

A Deus, por me dar forças para superar os obstáculos e motivação para desenvolver este trabalho de conclusão de curso.

Aos meus pais, que me ensinaram a nunca desistir dos meus objetivos e que sempre me incentivaram a estudar e persistir para atingir os objetivos.

Ao meu professor e orientador Edeilson Milhomem da Silva, que teve um papel fundamental no meu crescimento acadêmico e pelo apoio e incentivo no desenvolvimento deste trabalho. Aos demais professores do curso de Sistemas de Informação, agradeço a amizade que pude contar em diversos momentos e os ensinamentos.

A todos os demais colegas e amigos que, nos momentos de alegria ou nos momento de tristeza e abatimento, estiveram sempre dispostos a dar seu apoio.

RESUMO

BANDEIRA, Bruno Fernandes. **Desenvolvimento de um Mecanismo de Recomendação de Publicações Relacionados para a Plataforma *WordPress***. 2016. XX f. Trabalho de Conclusão de Curso (Graduação) – Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas (CEULP/ULBRA), Palmas/TO, 2016.

Sistemas de recomendação representam as preferências dos usuários com a finalidade de sugerir itens de interesse dos mesmos. Estes sistemas se tornaram fundamental por sua capacidade de filtrar informações que melhor se aplicam aos interesses do usuário de forma automática e efetiva. Para sugerir itens relacionados, o sistema de recomendação utiliza técnicas apropriadas ao contexto da aplicação, as quais processam as informações dos usuários e gera a recomendação personalizada. Diante disso, para entender o processo de funcionamento do sistema de recomendação, bem como, o mecanismo desenvolvido irá funcionar, este trabalho aborda os conceitos envolvidos em um sistema de recomendação utilizando os critérios para recomendação simples, e.g., palavras-chaves e técnicas de recomendação (Filtragem Baseada em Conteúdo). Com o objetivo de reduzir o tempo de processamento necessário à recomendação, é realizado o agrupamento das publicações mais similares utilizando a técnica de *clustering* K-means. Diante ao cenário apresentado, este trabalho tem como objetivo apresentar os conceitos e etapas envolvidas no desenvolvimento de um mecanismo de recomendação que possa ser implantando em um *plug-in* do *WordPress*, que irá agrupar as publicações mais similares e gerar recomendações. Esta ferramenta poderá ser agregada a uma aplicação e as recomendações das publicações deverão ser realizadas de forma *off-line*.

.

Palavras-chave: Filtragem. Sistemas de Recomendação. Recomendação Baseada em Conteúdo, *Clustering*.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DBSCAN	*Density-based Spatial Clustering of Applications with Noise*
FBC	Filtragem Baseada em Conteúdo
FC	Filtragem Colaborativa
FH	Filtragem Híbrida
IDF	Frequência Inversa do Termo
RI	Recuperação da Informação
SRI	Sistemas de Recuperação da Informação
SR	Sistema de Recomendação
TIC	Tecnologias da Informação e Comunicação
TF	Termos Frequentes

LISTA DE SÍMBOLOS

$TJ_{i,j}$	Frequência Inversa
idf_i	Frequência Inversa do Termo
$W_{i,j}$	Medida do Peso

1. INTRODUÇÃO

O termo “sociedade da informação” consolidou-se com os avanços das tecnologias de informação e comunicação e, dentro desse cenário, levou ao desenvolvimento de mecanismos e ferramentas que alteraram as formas de acesso e distribuição da informação.

Não há como negar que a internet é hoje o maior acervo de informação do mundo e encontra-se em crescimento constante. Neste cenário, a quantidade de informação disponível cresce muito além da capacidade de processá-la (MEDEIROS, 2013, p.1). Localizar o que se deseja em um ambiente de caráter interativo e dinâmico, entre tantas as opções, vem se tornando uma tarefa cada vez mais difícil para os usuários da internet. Os Sistemas de Recomendação surgiram em respostas a este problema, por exemplo, um Sistema de Recomendação de um *site* recomenda textos que os usuários gostariam de ler, músicas que possam interessá-lo, etc. Os resultados dessas recomendações são computados a partir da Recuperação da Informação.

O processo de Recuperação da Informação (RI) “é a área de pesquisa que se preocupa com a estrutura, análise, organização, armazenamento, recuperação e busca de informação” (Salton, 1968). A estrutura, análise, organização, e o armazenamento devem fornecer ao usuário acesso simples e objetivo, conforme suas preferências. O termo RI refere-se a mecanismos que possibilitam identificar facilmente e em tempo hábil objetos em um sistema, uma vez que o objeto armazenado tenha relação com as informações fornecidas pelo usuário. Neste contexto, aplicam-se os Sistemas de Recuperação de Informação (SRI), responsáveis pela representação, pelo armazenamento, pela organização e pelo acesso aos itens de informação (BAEZA-YATES & RIBEIRO-NETO, 1999, p.9).

Em contrapartida, pelo fato de os SRI retornarem um número elevado de resultados na busca, geram uma tarefa árdua para os usuários selecionarem a informação útil, uma vez que os mesmos devem realizar uma filtragem dos resultados da busca para encontrar os dados relevantes. Uma alternativa para auxiliar esse processo é o uso de Sistemas de Recomendação (SR). Um dos maiores desafios da área de SR é produzir recomendações que refletem na melhoria

da qualidade das recomendações realizadas, independentemente da manipulação de uma quantidade enorme de dados e das condições adversas que os mesmos se encontram. Existem três técnicas clássicas de recomendação, são elas: Recomendação Baseada em Conteúdo, Recomendação Colaborativa e Recomendação Híbrida.

Na técnica de Recomendação Baseada em Conteúdo (CORREIA, 2011) (BEZERRA, 2006) as preferências dos usuários são aprendidas com base em características específicas dos itens que este classificou ou que simplesmente visitou, e então o sistema cria um perfil dos conteúdos dos usuários. De posse do perfil do usuário e com base nos itens dos conteúdos, o sistema de recomendação utiliza aprendizagem computacional para selecionar os itens similares e recomendar. Na técnica de Recomendação Colaborativa (COSTA,2013) (MEDEIROS,2013) (VENSON, 2002) os itens são filtrados para um usuário baseando-se em experiências de outros usuários com gostos similares, assumindo a ideia de que pessoas com mesmo gosto possuem também os mesmos interesses (CORREIA, 2011, p.16). Já a técnica de Recomendação Híbrida (BORGES, 2010) combina as técnicas de Recuperação Baseada em Conteúdo e Colaborativa, "tendo como objetivo superar as limitações individuais de cada uma das técnicas e potencializar os seus benefícios" (MEDEIROS, 2013, p.8). Neste trabalho, foi contemplada apenas a técnica de Recomendação Baseada em Conteúdo.

Os Sistemas de Recomendação podem ser implementados através de técnicas que especifiquem classificação ou agrupamento de dados (ou *clustering*) (TAVARES, 2012) (FONSECA, 2010) (WIVES, 2004) (SILLA, 2002) (JAIN, 1999). As técnicas de agrupamento funcionam através da identificação de grupos de usuário que apresentam preferências semelhantes. O algoritmo K-Means, devido a sua simplicidade, tornou-se o mais utilizado no contexto de problemas relacionados a agrupamento de dados, servindo de parâmetro para funcionamento de outros algoritmos de agrupamento que surgiram posteriormente, como o algoritmo *Bisecting* K-Means. Outras técnicas surgiram baseadas na densidade (*Density-based*), em que definem clusters criando relações entre os objetos que se encontram em grupos com limite mais irregulares. No entanto, algoritmos com um único classificador como parâmetro, controlados pelo usuário (K *Nearest Neighbors*) podem obter melhores resultados, variando de problema para problema. Todavia,

após a identificação de grupos de usuários, a recomendação será direcionada mais fortemente para o usuário que se enquadra em grupos similares.

O *WordPress* é uma plataforma utilizada para publicação e Gerenciamento de Conteúdo na web, "sendo hoje a maior plataforma de Gerenciamento de Conteúdo do Mundo, com quase 70% do mercado" (WORDPRESS, Online). Diante desta perspectiva, este trabalho tem como objetivo definir um mecanismo de recomendação baseado em conteúdo, que possa ser implantado em um *plug-in* do *WordPress* que identifique e recomende posts relacionados.

A utilização de técnicas baseadas em conteúdo no desenvolvimento de um mecanismo de recomendação de publicações que possa ser implantado na plataforma *WordPress* pode vir a recomendar publicações de forma *off-line* que estejam mais próximas as necessidades dos usuário. Para isso, técnicas de *clustering* podem ser aplicadas para identificar grupos de publicações que apresentam preferências semelhantes e então a recomendação poderá ser direcionada mais fortemente para os usuários em que se a publicação visualizada se enquadra dentro dos mesmos grupos de similaridades.

Nas próximas seções serão apresentados os conceitos necessários para o desenvolvimento do trabalho como: referencial teórico, abordando os principais conceitos de Recuperação da Informação, Sistemas de Recomendação, Técnicas de Recomendação e Abordagens para Recomendação Baseado em Conteúdo; Materiais e Métodos utilizados; Resultados e Discussões obtidos, e, por fim, as Considerações Finais, abordando também possíveis melhorias e sugestões como Trabalhos Futuros.

2. REFERENCIAL TEÓRICO

Nessa seção são apresentados os conceitos necessários para prover o conhecimento teórico sobre a ferramenta desenvolvida. Para tanto, é necessário entender sobre Recuperação da Informação, Sistemas de Recomendação e Técnicas para Recuperação da Informação. O entendimento destes conceitos é premissa necessária para o desenvolvimento do projeto proposto neste trabalho, que é o desenvolvimento de um mecanismo de recomendação baseado em conteúdo.

2.1 Recuperação da Informação

A recuperação de informação (RI) faz parte da Ciência da Computação e lida como recuperar dados, informações ou documentos de alguma fonte de dados, sejam tabelas, documentos ou páginas na internet. Baeza-Yates e Ribeiro-Neto (1999, p.1) tratam a RI "como um processo que se inicia com a representação e armazenamento e estende-se até a organização e acesso à informação". Ou seja, a concepção de um sistema de RI se inicia com a definição da fonte de informação e as operações que podem ser executadas durante um processo de busca. Em seguida, é definida a visão lógica dos documentos, que viabilize tais operações e construção dos índices que armazene os termos, contidos nos documentos. Dessa forma, o processo de recuperação pode ser inicializado, com o usuário descrevendo a sua necessidade por meio de uma consulta. O RI interpreta a consulta e uma lista de documentos são ordenadas e apresentadas ao usuário.

O processo de recuperação tradicionalmente se dá pela necessidade de informação de um individuo. Dos sistemas usados em bibliotecas aos contemporâneos mecanismos de busca na web, ambos são inicializados à medida que o usuário formule e submeta sua consulta. A etapa posterior a isso, muito embora seja em um ambiente computacional, continua sendo atribuída a profissionais da informação (GIORDANO, 2011, p.32). Entretanto, com os avanços nas Tecnologias da Informação e Comunicação (TICs), a disseminação do uso da web e as constantes melhorias nos mecanismos de buscas, tarefas antes desempenhadas por profissionais habilitados passaram a ser realizadas corriqueiramente por usuários finais.

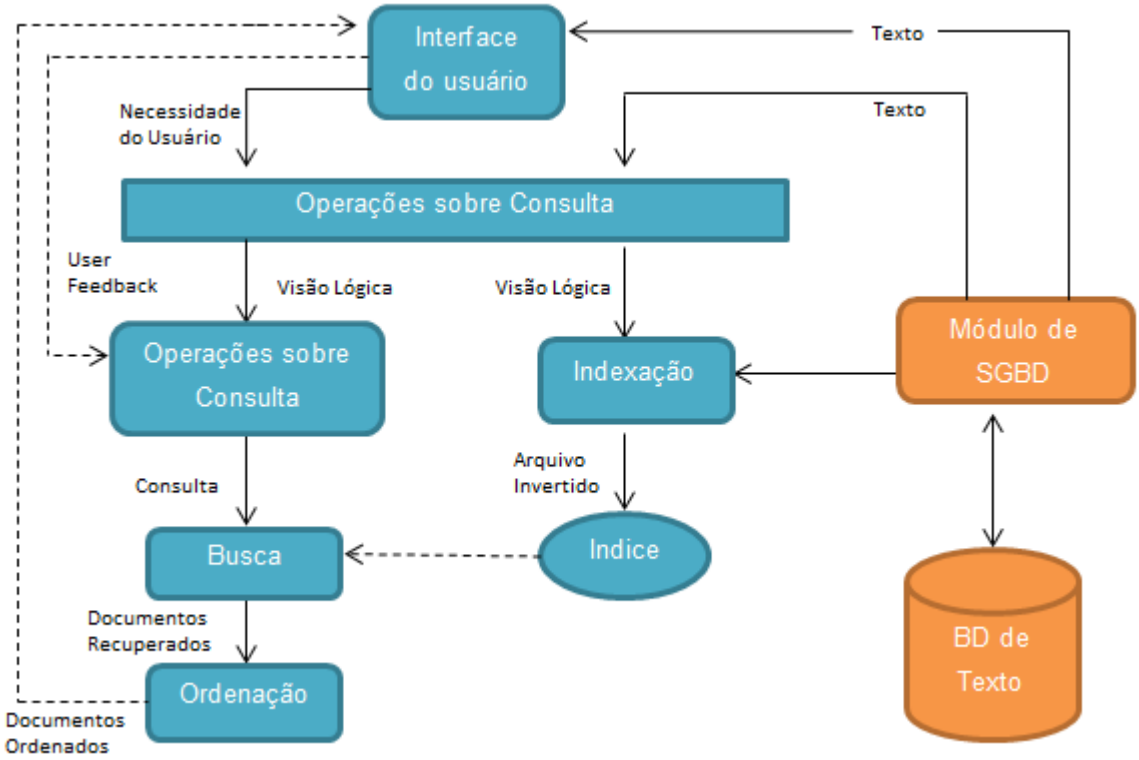
Com os avanços na área da recuperação da informação o processo de busca tornou-se corriqueiro, conforme descreve Saracevic (1999, p.162) “todo mundo é “buscador” hoje em dia”. Todavia, a recuperação da informação se propagou por várias áreas, uma vez que, antes se concentrava apenas em formato textos, originalmente, “existem pesquisas e esforços pragmáticos devotados à recuperação da informação em música, vídeo, fotografias e imagens em movimento e multimídia” (SARACEVIC, 2010, p.162).

O termo Recuperação da Informação foi criado por Calvin Mooers, sendo considerado um campo de pesquisa interdisciplinar, baseadas em muitas áreas. Dos cartões perfurados aos sistemas com múltiplas possibilidades de formação de busca, o processo de RI foi evoluindo e se adequando ao contexto, na medida com a tecnologia evoluía. Entretanto, com a criação da web e "em meio a todos esses avanços, cresciam também as bases de dados" (GIORDANO, 2011, p.35). A popularização da internet e o surgimento de busca on-line por usuários finais para tratar essa maciça quantidade de informação existente, incentivaram estudos na área de RI no ciberespaço, representando significativa evolução no desenvolvimento dos sistemas de informação.

2.1.2 Sistemas de Recuperação da Informação

Os Sistemas de Recuperação da Informação (SRI) são uma forma de recuperar automaticamente informações relevantes a partir de uma determinada consulta. Essas informações podem ser documentos, dados que descrevem documentos, dados armazenados em bases de dados relacionais e documentos hipertexto (texto, som e imagem) presentes, por exemplo, na Internet ou Intranet- (DEPPLER et al, 2005, p. 4). A **Error! Reference source not found.** apresenta o processo de recuperação de informação comum à maioria dos sistemas.

Figura 1. Processo de Recuperação de Informação



Fonte: Adaptada de OLIVEIRA, 2005, p 24.

Na Figura 1, a arquitetura representa um processo de RI e embora não esteja explicitamente visível no esquema apresentado, o processo de RI engloba todas as etapas referentes ao processo de consulta, ordenação dos documentos e o processo de indexação. Conforme Figura 1, arquitetura é composta basicamente por: Interface do Usuário-, onde o usuário apresenta sua necessidade de pesquisa; Operações sobre textos, como, por exemplo, a extração de palavras que não são úteis para a pesquisa-; Operações sobre Consultas-, responsáveis pela formulação dos termos relevantes para consulta -; essas operações são descritas por meio de expressões booleanas usando conectores lógicos: *AND*, *OR* e *NOT*-; a Busca, que retorna uma lista de documentos; a Ordenação-, responsável pelo ranking (ordenação) dos documentos recuperados; e a Indexação-, responsável pela criação dos termos que melhor identificam um terminado documento, para que possam ser construídas as estruturas sobre o quais o processo de consulta realizará as pesquisas.

Para Baeza-Yates & Ribeiro-Neto (1999, p.21) existem duas formas de realizar o processo da RI: ad-hoc e com filtragem. No forma ad-hoc, são realizadas consultas em uma base com a coleção de documentos e é retornado o resultado. A ad-hoc leva em consideração a consultado realizada, independente do usuário. Em contrapartida, na recuperação por filtragem, faz uso do perfil do usuário, retornando um resultado utilizando a descrição das preferências contidas no perfil do usuário com a descrição dos itens que estão na base de dados.

Há uma diversidade de modelos de RI que oferecem diferentes especificações que podem ser utilizadas em adequação às informações armazenadas e podem ser classificados de acordo com a forma como implementam ou não os processos de indexação, operações sobre consulta, busca e ordenação. Conforme explicam Baeza-Yates e Ribeiro-Neto (1999, p.17), um modelo de RI é composto de: (a) um conjunto de visões lógicas, ou representações, de documentos em uma coleção; (b) um conjunto de visões lógicas, ou representações, da informação requerida por usuários; (c) um framework para modelar documentos, consultas e suas relações; e (d) uma função de ordenação que associa um número com uma consulta e um documento.

GRENNGRASS (2000, p.13) aborda duas categorias principais de modelos de RI: o modelo semântico, que implementam análise semântica e sintática na descrição dos documentos em linguagem natural; e o modelo estatístico (ou clássico), que atribui medidas estatísticas que se referem a comparações entre uma consulta e um documento.

Segundo Baeza-Yates & Ribeiro-Neto (1999, p. 24) os modelos clássicos consideram que cada documento é descrito como um conjunto de palavras-chave denominadas termos (ou índices). Um termo é uma palavra que resume o conteúdo de um documento. Para cada termo, pode-se um valor numérico (peso) que indicará o grau de relevância daquele termo ao descrever o conteúdo de um documento.

Estão incluídos nesta categoria os modelos booleanos, vetorial e probabilístico, conforme Figura 2.

Figura 2. Divisões do Modelo Clássico

Fonte: Tsuji, 2008, p.17

Modelo Booleano - baseado na teoria dos conjuntos, faz uso de operadores lógicos na consulta, para recuperar os documentos, podendo essa consulta ser formada por elementos lógicos como AND, OR e/ou NOT. Os operadores lógicos indicam apenas se o termo está ou não presente no documento. Para Tsuji (2008, p.17) pertencem a essa modelo, Lógica *Fuzzy* e Booleano *Extendido*.

Modelo Probabilístico - é considerada uma aplicação direta da teoria das probabilidades, na qual utiliza pesos binários para representar os documentos, onde determina a presença ou ausência de termos. A partir da consulta do usuário, há um conjunto de documentos que possui documentos relevantes e não relevantes para o usuário (STOKOVIAK, 2011, p.15). Conforme Figura 2, entram nesse modelo Redes de Inferência e *Belief Networks*.

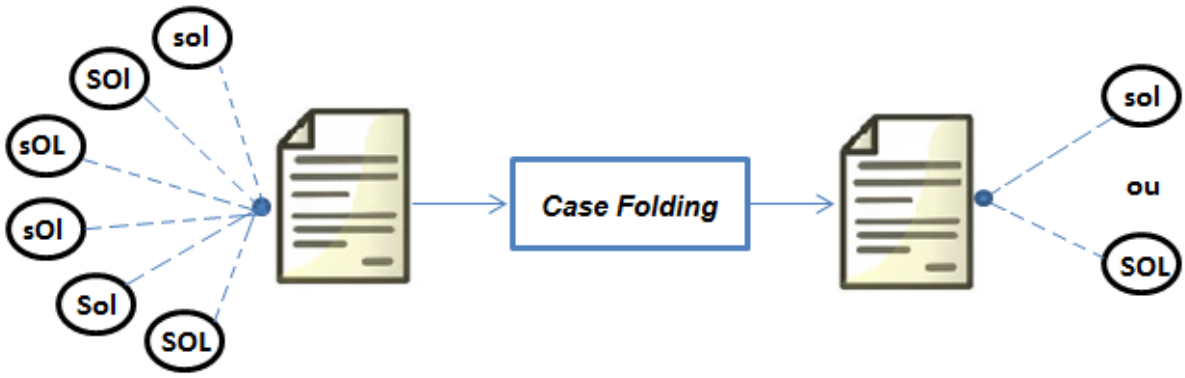
Modelo Vetorial - considerado um modelo algébrico, fazendo uso de vetores n-dimensionais para representar os documentos, em que n indica a quantidade de termos únicos que ocorrem no interior de todos os documentos (STOKOVIAK, 2011,

p.15). Por consequente, busca-se encontrar os vetores que mais se aproximam do vetor equivalente à consulta submetida, ordenando o resultado conforme o grau de relevância do documento. Conforme Figura 2, incluem nesse modelo o Vetor Generalizado, Índices Semânticos e Redes Neurais.

O cálculo do peso de cada termo calcula a similaridade entre os documentos da base e a consulta. No entanto, antes que ocorra o cálculo do peso de cada termo, para analisar os vetores que mais se aproximam do vetor equivalente, um documento pode passar por um processo de Recuperação da Informação. A etapa inicial do processo de Recuperação da Informação é chamada de Pré-Processamento ou Preparação dos Documentos. Esta etapa é responsável por determinar quais termos do documento descrevem melhor o seu conteúdo, com o propósito de diminuir a complexidade da representação deste conteúdo. Para realizar essa etapa são destacadas algumas técnicas, como: *Case Folding*, *Stop Words* e *Stemming*. A seguir são explicadas mais detalhadamente as três técnicas citadas:

A técnica de *Case Folding* trata da padronização dos caracteres de um documento. Esta padronização pode ser todo o texto em letras maiúsculas ou em letras minúsculas. A Figura 3 apresenta um exemplo de aplicação do *case folding*.

Figura 3. Exemplo de Case Folding.

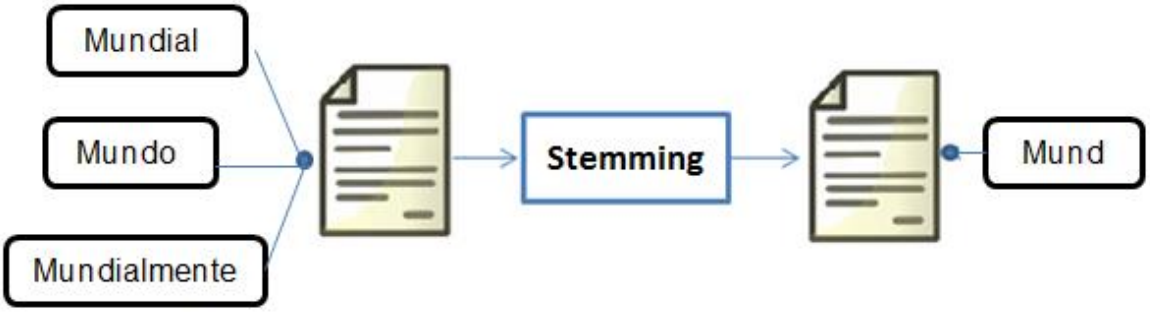


O processo de *case folding* padroniza as palavras para que estas possam ser processadas como uma única palavra, fazendo com que o processo de comparação entre os caracteres seja facilitado. Conforme demonstra a Figura 3, ao aplicar o *case folding* nas palavras "sol", "SOl", "sOL", "sOl", "Sol" e "SOL", todas serão

transformadas para o formato comum em letras maiúsculas ("SOL") ou em letras minúsculas ("sol").

A técnica “Stemming é o processo de converter cada palavra para o seu radical, eliminando sufixos representados por flexões verbais e plurais” (SILLA & KAESTNER, 2002, p. 1-2). Isso é, cada variação morfológica de um termo é eliminada e considera-se apenas a sua raiz, eliminado os prefixos e/ou sufixos. Realizando essa etapa, a quantidade de palavras diferentes a serem tratadas no texto será reduzida. A Figura 4 apresenta um exemplo de aplicação do processo *stemming*.

Figura 4. Exemplo de Stemming.



A Figura 4 apresenta um exemplo de aplicação da técnica de stemming, na qual as palavras "Mundial", "Mundo" e "Mundialmente" tiveram seus respectivos sufixos "ial", "o" e "ialmente" removidos, resultando em um mesmo radical "Mund".

Stop Words é o processo responsável por eliminar itens que comumente não trazem muito significado para o tema do documento, como pronomes, artigos, preposição e conjunção, e aparecem com muita frequência. Segundo Salton & McGill (1983, p.30) uma lista de *stop words* é uma lista de palavras que não possuem relevância para o documento e realiza a remoção de 40 a 50% do total de palavras de um texto. A eliminação desses itens não será prejudicial para a Análise da Similaridade entre os textos, pois esses termos servem apenas para dar sentido gramatical ao texto.

A Tabela 1 demonstra um exemplo da aplicação dessa técnica.

Tabela 1. Exemplo de *Stop Words*

Texto do Documento	***Stopwords***	**Termos Relevantes**
Brasileiro passa mais tempo na internet *que* na TV, diz pesquisa.	Mais	Brasileiro
	Na	Passa
	Que	Tempo
		Internet
		TV
		Diz
		Pesquisa

No exemplo demonstrado na Tabela 1, é aplicada a técnica de *stop words* que retira do documento as palavras sem significado relevante. Na Tabela 1, as palavras retiradas foram “mais”, ”na”, “que” deixando somente os substantivos e verbos.

Uma vez que a preparação do documento tenha sido concluída, é possível obter um documento contendo apenas os termos mais relevantes e a partir desse momento, seguir para o cálculo do peso de cada termo e então obter a medida do peso.

O vetor TF-IDF serve para obter a medida dos pesos dos itens no vetor (BORGES, 2011, p.20). No entanto, para calcular o TF-IDF, é necessário realizar o cálculo da TF (*Term Frequency* ou Frequência do Termo) e IDF (*Inverse Document Frequency* ou Frequência Inversa). A formalização do cálculo do TF pode ser representada na equação matemática:

$$TJ_{i,j} = \frac{\text{freq}}{\max f\ req}$$

Na equação acima, a variável *freq* representa a frequência da ocorrência de determinado termo em um documento, sendo dividida pela frequência do termo que mais ocorre no documento (*maxfreq*). Por exemplo, se existe um termo t1 cuja frequência é 8, um termo t2 com frequência 7 e um termo t3 com frequência 5, para

saber o peso do termo t2, basta dividir a frequência de t2 pela frequência de t1, uma vez que reflete o termo mais frequente.

Já a frequência inversa do termo (IDF), tem como objetivo identificar os termos que ocorrem com frequência em um documento e também aparecem frequente em outros documentos de temas diferentes, denominados termos comuns. A formalização do cálculo da frequência inversa pode ser representada através da fórmula:

$$idfi = \log \frac{N}{\mathrm{n}i}$$

A equação apresentada acima, leva em consideração a quantidade total de documentos existentes (N) e a quantidade de documentos que ocorrem o determinado termo ($\mathrm{n}\, i$). Portanto, se existem 10 documentos e em 6 documentos ocorre o termo t1, assim para encontrar o IDF basta calcular o logaritmo da divisão da quantidade total de documentos existentes (10), pela quantidade de documentos que ocorre o termo t1 (6).

Após o cálculo da TF e IDF, é possível obter a medida do peso. A medida do peso é obtida através da multiplicação da TF pela IDF, conforme apresentada na fórmula:

$$W\mathrm{i,j} = \mathrm{tf}\, x\, \mathrm{idf}$$

Depois de medido o peso de cada termo de cada documento, é possível calcular a similaridade entre os documentos. Concluído o processo para o cálculo de similaridade entre os documentos, é possível montar uma lista contendo os documentos mais similares. A próxima seção abordará sobre os Sistemas de Recomendação.

2.2 Sistemas de Recomendação

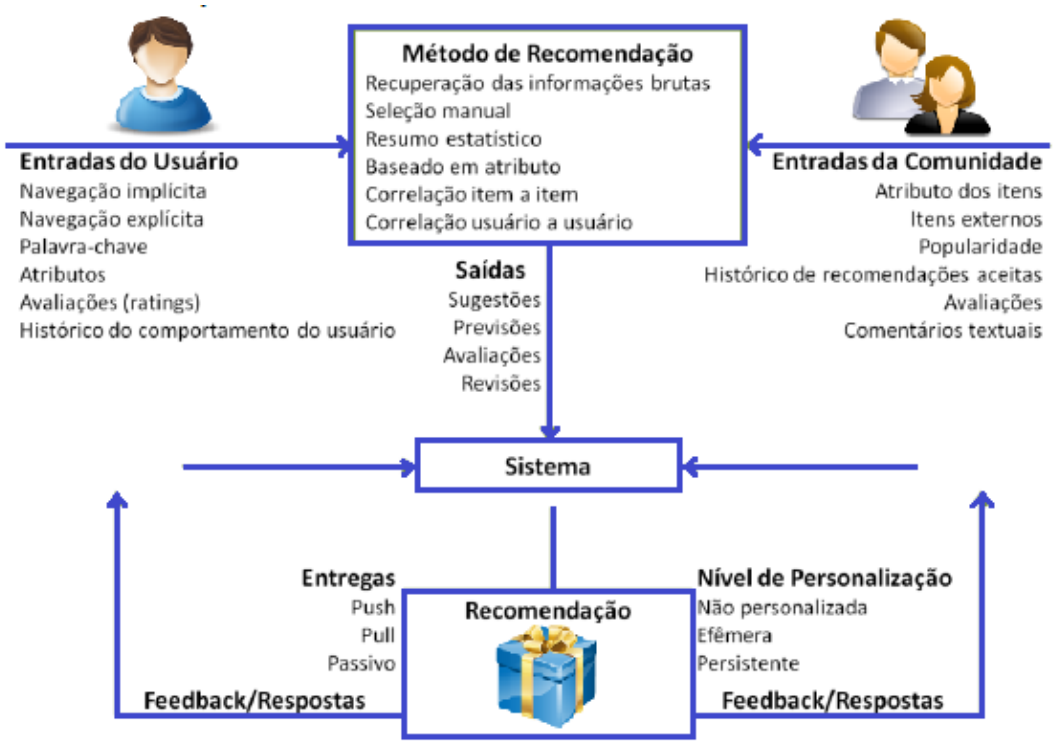
Com o advento das Tecnologias de Informação e Comunicação cresce a quantidade de informação e o acesso facilitado das mesmas por intermédio da Internet. Em contrapartida, as pessoas se deparam com uma diversidade de opções e muitas vezes, indivíduos possuem pouco ou quase nenhuma experiência para selecionar conteúdo relevante dentre as opções apresentadas. Para minimizar as dúvidas

frente às escolha entre alternativas, geralmente se confiam nas recomendações que são passadas por outras pessoas, ou através de mecanismo automáticos, opiniões adversas, dentre outros.

Para Borges (2010, p.14) "pode se dizer que Sistemas de Recomendações são utilizados para auxiliar os usuários a identificarem serviços ou produtos de interesse que estejam dentro de uma grande quantidade de opções". Dessa forma, um SR auxilia no aumento da capacidade e eficácia dos sistemas em disponibilizar conteúdo relevante para o usuário, em meio a opções existentes. Venson (2002, p.16) estende essa definição, onde um sistema de recomendação é "um mecanismo capaz de aprender através de iterações dos usuários com o sistema, a fim de obter experiências para poder recomendar, dentro os produtos disponíveis", ou seja, o que tiver uma maior relação com o usuário. Assim, os SRs podem identificar características semelhantes aos usuários baseado em perfis, e recomendar um produto ou serviço de interesse, que não seja diretamente especificado.

A Figura 5 apresenta as etapas de funcionamento de um SR:

Figura 5. Etapas de um Sistema de Recomendação



Fonte: Adaptada em BORGES, 2011, p. 12.

Conforme apresentado na Figura 5, um SR recebe como entrada informações sobre o usuário; com sua forma de navegação, informações sobre seu comportamento e sua participação na comunidade do site, em seguida, utiliza os métodos de recomendação para gerar sua saída (recomendações) e por fim, são apresentadas as recomendações aos usuários.

Conforme apresentado na Figura 5, os SRs podem possuir algumas estratégias ou grau de personalização para realizar recomendação: não personalizada, efêmera e persistente.

O SR não personalizado são aqueles que utilizam os mesmos critérios de recomendação para todos os usuários. Por exemplo, em um site em *WordPress*, a cada nova publicação criada, o mesmo é recomendado para todos os usuários que visitam o site, bem como o último cadastrado (mais recente) ou uma publicação aleatório. Na Figura 6 é apresentada uma exemplificação da recomendação não personalizada.

Figura 6. Recomendação Não-Personalizada

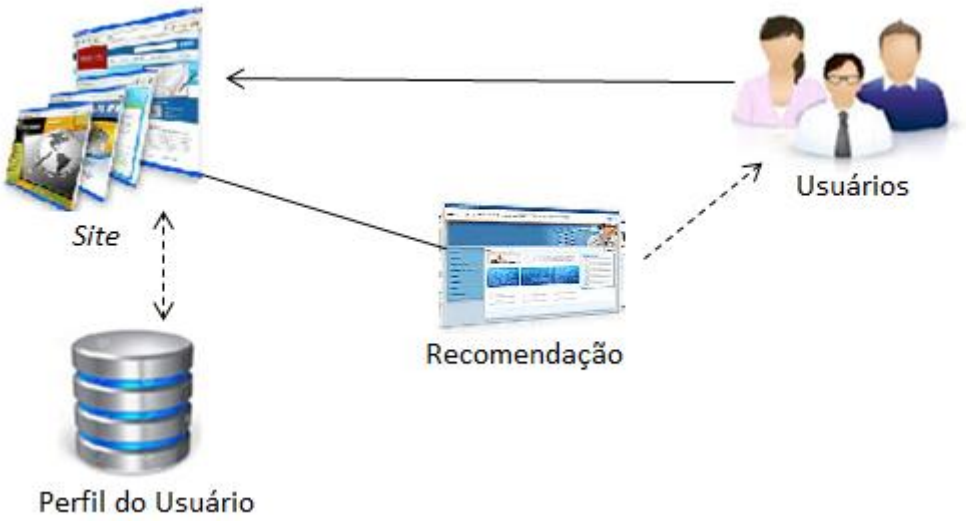
A Figura 6 representa um ambiente de recomendação não personalizado de publicações em um site em *WordPress*. Neste cenário, a cada nova publicação, o mesmo é recomendado para todos os usuários visitantes. Como não há uma personalização nas recomendações, para todos os usuários visitantes é gerada uma recomendação das últimas publicações disponibilizadas.

No SR efêmero o mecanismo recomenda na medida em que o usuário está acessando o sistema, por base nas informações coletas pelo sistema. Por exemplo, quando um usuário realiza uma pesquisa por *posts* no mecanismo de busca do site

ou o mecanismo analisa os posts que ele visitou, o sistema utiliza estas informações para gerar recomendações para o usuário.

No SR persistente o sistema utiliza as informações armazenadas no perfil dos usuários, desde informações de históricos e preferências pré-definidas. A Figura 7 apresenta uma exemplificação de recomendação persistente.

Figura 7. Recomendação Persistente



Na Figura 7, as preferências dos usuários estão armazenadas em seus perfis. Essas preferências podem ser salvas ao ser realizadas o cadastro do usuário no site ou mesmo por meio de técnicas de identificação automática. Na próxima seção é descrito sobre a construção de perfis dos usuários.

2.2.1 Perfil de Usuários

O perfil de usuário é um conjunto de dados pessoais associados a um usuário específico. O perfil do usuário reúne características que representam o usuário em aspectos, tais como suas preferências e interesses pessoais. Conforme Rousseau et al (2004, p.38) fisicamente, o perfil do usuário pode ser visto como uma base de dados onde a informação sobre o usuário é armazenada e pode ser dinamicamente mantida.

Para Duved et al (2008 apud CASTRO & BARBOSA, 2009, P. 17) três importantes etapas são necessárias para a construção do perfil do usuário, que servirão como base para as recomendações, que são:

- Modelagem do usuário: momento em que se definem quais os tipos de características do usuário podem ser considerados importantes, ou seja, quais os interesses e preferências do usuário. Essas características são classificadas em críticas ou não para o sistema e separadas em grupos.
- Perfil do usuário: características que descrevem o usuário são armazenadas para compor o perfil do usuário. As características são obtidas por meio das interações do usuário no sistema ou informações diretamente cadastradas pelo usuário. Dessa forma, esta etapa consiste na formação da base de interesse do usuário.
- Personalização: consiste na etapa final, na qual o perfil do usuário é utilizado para realizar as recomendações.

Para representar um perfil do usuário, é preciso coletar informações a respeito das preferências do usuário, para assim, serem geradas as recomendações com base no perfil criado, respectivo a cada usuário. A etapa inicial para o levantamento das informações do perfil consiste na coleta das informações (coleta explícita e coleta implícita) e consequentemente na manutenção deste perfil e que será abordada na próxima seção.

2.2.1.2 Geração e Manutenção de Perfil do Usuário

Para que o sistema consiga realizar recomendações personalizadas, requer que o sistema possa identificar o usuário no momento em que este acessa o sistema. Duas formas habituais de identificação de usuários são:

- Identificação no Servidor: eventualmente, nesta forma de identificação é disponibilizada ao usuário uma área de cadastro com informações pessoais, assim como um usuário único e senha, que ficaram armazenadas em um banco de dados no servidor. Posteriormente, ao acessar o sistema, o usuário irá se autenticar no servidor com seu usuário único e senha e, dessa forma, o website identificará o usuário que nele se conecta com maior precisão. Na figura 9 é apresentado o mecanismo de autenticação nos serviços do Google, na qual com uma única conta, é possível acessar todos os serviços oferecidos (*E-Mail*, *Youtube*, *Driver*, e outros).

Figura 8. Identificação no Servidor (Serviços do Google)

- Identificação no Cliente: para este recurso, normalmente utiliza-se os *cookies*, em que consiste em um mecanismo pelo qual o website consegue identificar que determinado computador está conectado, haja vista que o mesmo já tenha acessado o website em um período anterior. No entanto, uma desvantagem deste método, é que o mesmo assume que a máquina conectada é utilizada sempre pela mesma pessoa, desprezando a possibilidade de que várias pessoas possam utilizar um único computador para acessar o website. Trata-se de um mecanismo mais simples do que a identificação através do servidor, porém bem menos confiável. Na Figura 9, apresenta esse tipo de situação. Como pode ser observado no canto superior esquerdo, ao acessar a página do site, o nome de identificação do usuário é carregado automaticamente e questionado o usuário se trata do mesmo usuário apresentado.

Figura 9. Identificação no Cliente (cookies)

Após a identificação do usuário, é possível coletar dados sobre formas implícitas ou explicitas, permitindo então, a geração e manutenção de seu perfil. Na coleta explícita o usuário informa espontaneamente o que lhe é importante, demonstrando de alguma maneira se gosta ou não de determinado produto, categoria, etc., fornecendo notas, comentário, avaliações, marcando em uma *checkbox* etc. No exemplo a seguir (figura 10), o usuário define as seções que consideram favoritas, marcando-as.

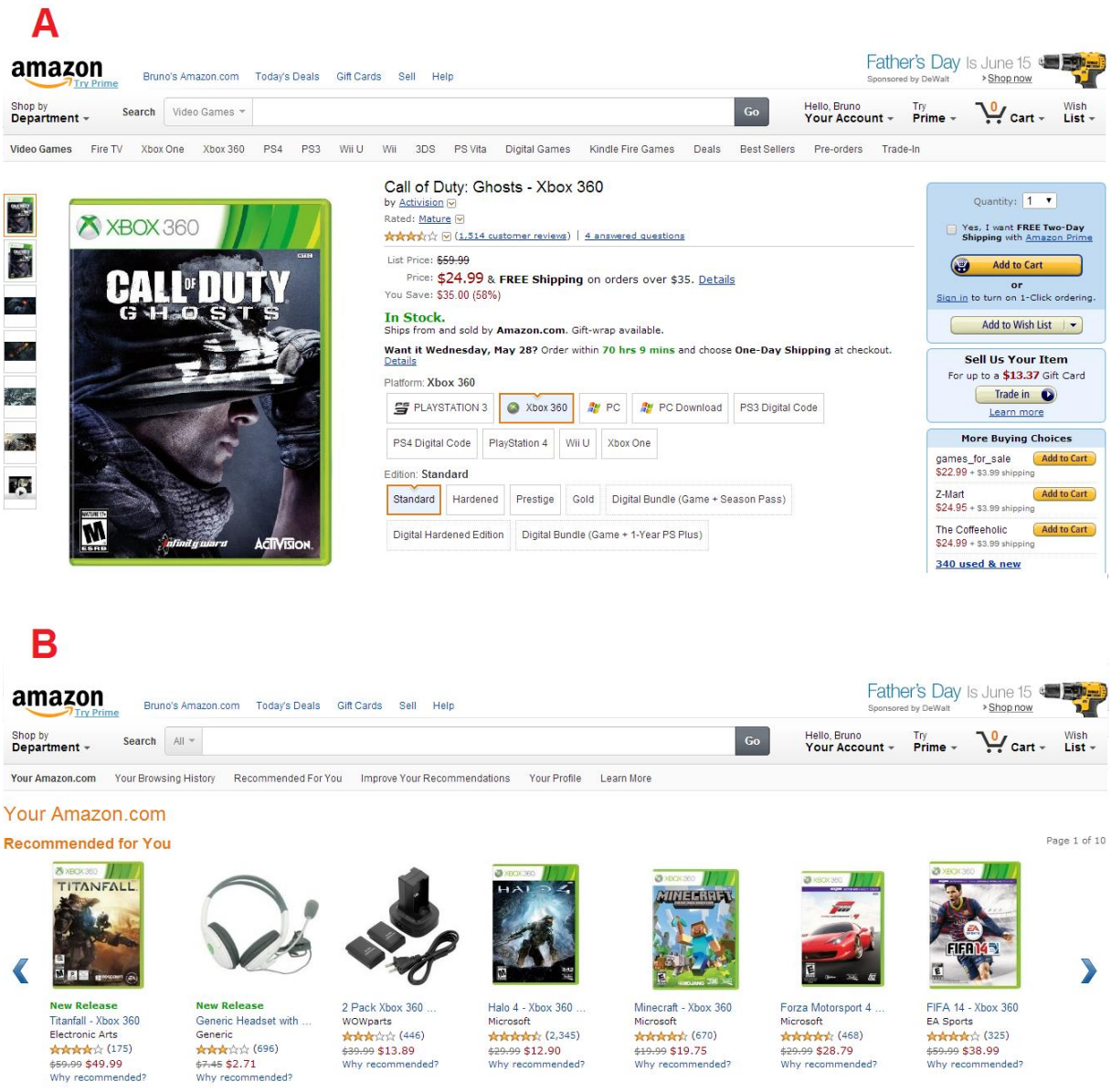
Figura 10. Coleta Explícita

No exemplo apresentado pela figura 10, o usuário autenticado no sistema, tem a opção de marcar as áreas de interesse dele no que diz respeito a categorias disponíveis no website. Por considerar o fato de o próprio usuário fornecer as informações ao sistema, este tipo de coleta é considerado mais confiável, no entanto, nem sempre o usuário estará disposto a colaborar com informações para o sistema, ou mesmo disponibilidade para isso.

Na coleta implícita, por meio das ações do usuário no sistema, ou seja, através do monitoramento do comportamento e da interação do usuário no *website* (por exemplo, durante a navegação no site, através do histórico de aquisição de um produto, páginas visitadas, principais palavras chaves buscadas, entre outros)- é possível inferir informações sobre suas necessidades e preferências. Essa técnica tem como propósito, conhecer melhor as preferências dos usuários sem que eles forneçam informações explicitamente para o sistema. A Figura 11 (A) apresenta a captura de informações de forma implícita no website da Amazon.com, sem que o mesmo perceba. Isso é possível através das interações (no caso em particular, as visitas aos produtos) do usuário no sistema. Na Figura 11 (B), é apresentada uma

página da Amazon.com personalizada para um usuário que se mostra interessado por acessório e jogos do console *X-Box* 360, inferida pelo sistema por visitas do usuário (Figura 11 A) em outros produtos semelhantes.

Figura 11. Coleta e Página Personalizada a partir de Interesses Implícitos



Com os dados dos usuários armazenados, são realizadas as comparações utilizando técnicas e algoritmos, para descobrir a similaridade entre os itens armazenados. Com o perfil dos usuários definidos, é possível identificar itens que sejam das preferências dos usuários ou mesmo o mais próximo possível e assim recomendar ao mesmo.

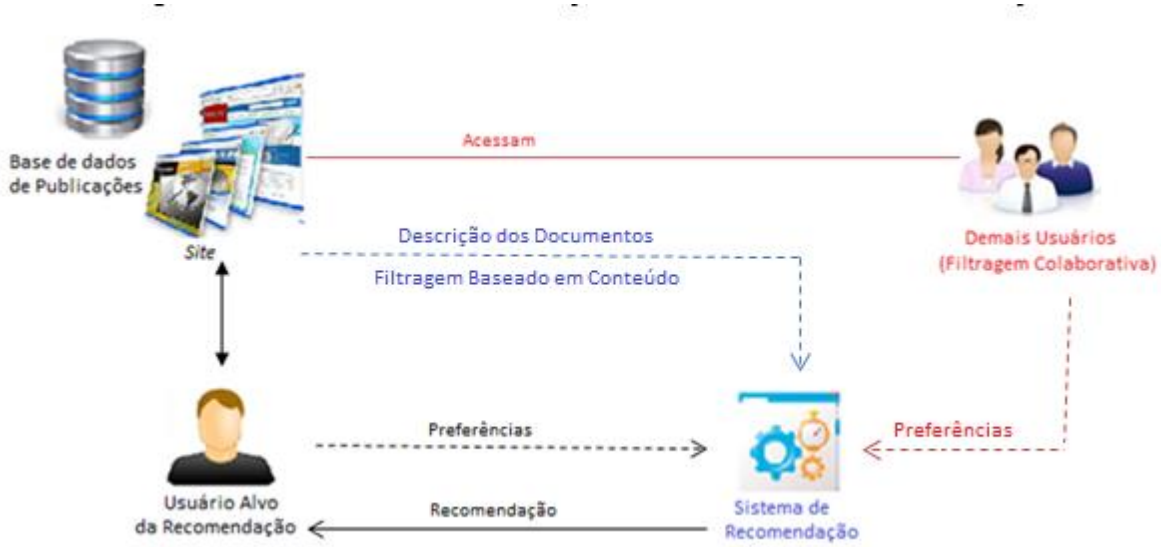
Nas seções subsequentes, são descritas as técnicas de classificação dos Sistemas de Recomendação, bem como detalhado as técnicas de recomendação baseado em conteúdo, foco deste trabalho.

2.3 Classificação dos Sistemas de Recomendação

Alguns dos grandes desafios na área de Sistemas de Recomendação trata do poder de processamento dos SR. A demanda dos dias atuais necessita que alguns mecanismos de recomendação manipulem milhares de dados em tempo real, como os SR destinados ao contexto de *e-commerce* (VENSON, 2002)(BORGES, 2010). Outro grande desafio lida na melhoria da qualidade das recomendações geradas. Dessa forma, uma grande variedade de técnicas e formas diferentes para se trabalhar de forma escalável e com melhorias na eficiência das recomendações são propostas para cada contexto. Os SRs são comumente classificados de acordo com a técnica apresentada, visando à identificação de padrões de comportamento e fundamentam o funcionamento dos SRs.

Costa et. al (2013, p.60), distingue cinco principais classes de técnicas de recomendação: Filtragem Colaborativa (FC), Filtragem Baseada em Conteúdo (FBC), Filtragem Demográfica, Filtragem Baseada em Utilidade e Filtragem Baseada em Conhecimento. As técnicas mais adotadas são baseadas na FC e FBC. A Figura 12 demonstra de forma genérica o funcionamento de um SR utilizando as técnicas de recomendação FBC e FC.

Figura 12. Técnicas de Recomendação em um Sistema de Recomendação



A Figura 12 apresenta um modelo genérico de um SR mostrando o processo realizado ao utilizar as técnicas de recomendação FBC e FC. Na Filtragem Baseada em Conteúdo é levado em consideração o perfil do usuário que se analisa e a base de dados (apresentada como uma base de publicações de um Portal) para fazer o relacionamento das publicações a serem recomendados baseados nos critérios selecionados ao perfil do usuário. A FBC também pode realizar a recomendação sem considerar o perfil do usuário, apenas analisando a similaridade entre as publicações existentes na base de dados. Em contrapartida, na Filtragem Colaborativa, são analisadas as preferências do usuário, bem como os demais usuários, para realizar uma comparação entre os perfis e observar quais perfis aos demais usuários são mais similares ao do Usuário Alvo.

Este trabalho tem como foco criar um mecanismo de recomendação por meio de técnicas de Filtragem Baseado em Conteúdo. Dessa forma, a próxima seção abordará apenas a FBC, na qual serão apresentados algoritmos inerentes a essa abordagem.

2.3.1 Filtragem Baseado em Conteúdo

A abordagem FBC deriva do trabalho de investigação e avanços na área da filtragem de informação e filtragem de conteúdo. A Filtragem Baseada em Conteúdo (FBC) parte do princípio de que usuários tendem a ter uma tendência natural a se interessar por itens semelhantes aos que demonstraram interesse anteriormente

(Stackoviak, 2001, p 11). Deste modo, vários itens são comparados com itens que foram avaliados anteriormente e os mais similares são recomendados para o usuário. Em contrapartida, para que haja essa recomendação, além de levar em consideração a análise de conteúdo do item, analisa-se o perfil do usuário.

O processo de recomendação baseado em conteúdo seleciona os itens mais similares aos itens identificados no perfil de interesse do usuário, que por sua vez, a construção desse perfil pode ser realizada de forma explícita, com uso de questionário, ou implícita, obtidas na coleta através do conteúdo dos itens que o usuário consumiu.

Para que seja possível estabelecer a similaridade entre os itens, faz-se necessário identificar atributos em comum entre os itens, para que possam ser comparados. Tais atributos correspondem às informações que o usuário consome ou forneceu para o sistema. Todavia, é possível que, em alguns casos, identificar os termos similares nos itens seja uma tarefa que exige alguma dificuldade.

Estes atributos são extraídos dos itens e compõem o perfil de cada item. O perfil dos itens é representado pelo vetor TF-IDF que contém a medida do peso de cada item de cada documento. Depois de medido o peso de cada termo de cada documento, é possível calcular a similaridade entre os documentos.

Assim como outras técnicas, a FBC possui suas vantagens e desvantagens (Borges, 2011, p.23)(Cazella, 2010, p.17):

- Vantagens
    - A utilização desta técnica com preferências incomuns entre usuários torna-se uma ótima alternativa, uma vez que se baseia nos itens já consumidos pelo usuário, e não necessita da contribuição de outros usuários com perfis parecidos;
    - Por levar em consideração apenas as preferências do usuário, as recomendações são mais precisas;
    - Todos os documentos da base de dados podem ser recomendados, levando em consideração suas similaridades ao perfil do usuário.
- Desvantagens

- A técnica de recomendação baseado em conteúdo não tem a capacidade de inovar em relação as recomendações, uma vez que essa técnica baseia-se somente nos itens que já foram consumidos.
- Recomendações imprecisas podem ser realizadas quando o usuário é novo e não possui muitos itens consumidos ou avaliados.
- As recomendações se limitam as características textuais dos atributos dos itens a serem recomendados.

Conforme mencionado, Sistemas de Recomendação baseados em conteúdo recomendam itens similares que o usuário aprovou no passado. Assim, vários itens são comparados com itens que foram avaliados positivamente e os mais similares serão recomendados. No entanto, este trabalho propõe-se a utilização de técnicas de análise e classificação prévia antes da aplicação de alguma técnica de recomendação à base de dados. A utilização de *clustering* tem como intuito comparar os resultados obtidos na aplicação de técnicas de recomendação, reduzindo o espaço de busca, criando *clusters* compostos por grupos de publicações que têm características similares. Dentre as técnicas de análise de dados mais utilizadas, a *clustering* tem se destacado.

2.4 Abordagem para Recomendação Baseada em Conteúdo

Geralmente a quantidade de informação a serem consideradas pelas técnicas de recomendação é tão grande que pode tornar o tempo de resposta um problema, ou seja, apresentando problema de escalabilidade. Uma das principais técnicas utilizadas para melhorar a escalabilidade em sistemas de recomendação é o *clustering* (também chamado de agrupamento). C*lustering* consiste no processo de agrupar documentos similares (ou relacionados) em categorias, denominadas por clusters (BAEZA-YATES & RIBEIRO-NETO, 1999 p.124) Segundo Oliveira (2005, p.43) “*clustering* tem por objetivo definir e identificar, de forma automática, grupos de objetos dentro de uma coleção”. Para criar *clusters*, são desenvolvidos métodos que utilizam um conjunto de informações agregadas aos documentos. Esses métodos são desenvolvidos levando em consideração características particulares do domínio onde o mesmo será aplicado.

C*lustering* compõe uma das tarefas de aprendizagem de máquina, e na literatura, existem dois tipos de algoritmos de aprendizagem de máquina: não-supervisionado e supervisionado. Utilizam-se algoritmos de aprendizagem de máquina supervisionada, quando se tem conhecimento prévio das classes do sistema. Para se definir as classes, o especialista no domínio deve definir um conjunto de exemplos contendo subconjunto como classes iniciais. Posteriormente, inicia-se a fase de treinamento, na qual é utilizado algoritmo/técnica de aprendizagem de máquina. Nessa fase, são geradas coleções de regras de classificação baseadas na análise do conjunto de treinamento fornecido, e assim, são redefinidos classes e os itens são classificados. Ao fornecer novos itens para a base de treinamento, é utilizado novamente algoritmo/técnicas de aprendizagem de máquina e as classes são redefinidas e os novos itens classificados (CORREIA, 2011, p.16). De maneira oposta a aprendizagem de máquina supervisionada, utiliza-se o algoritmo não supervisionados quando não se tem conhecimento prévio das classes do sistema. Nessa abordagem, o algoritmo generaliza os itens, sem a intervenção do usuário, baseando-se na observação e descobertas de padrões dos dados (CORREIA, 2011, p.18). Os elementos similares são agrupados (formação de *clustering*) em classes, que definem padrões para os demais itens, associando assim as classes do sistema as quais os itens futuros serão associados.

Nas seções seguintes são apresentados os algoritmos/técnicas mais utilizados na etapa de *clustering*.

2.4.1 K-Means

As técnicas de *clustering* foram desenvolvidas com o intuito de identificar e agrupar objetos. Os algoritmos de *clustering* são utilizados com muita frequência em aplicações que necessitem de busca por padrões, na possibilidade de criar conjuntos de objetos que de alguma forma contenham similaridades. Uma das principais técnicas de clusterização são os algoritmos particionais, que tem como característica a criação de aglomerados, fazendo assim, várias iterações (passagens) nesse conjunto.

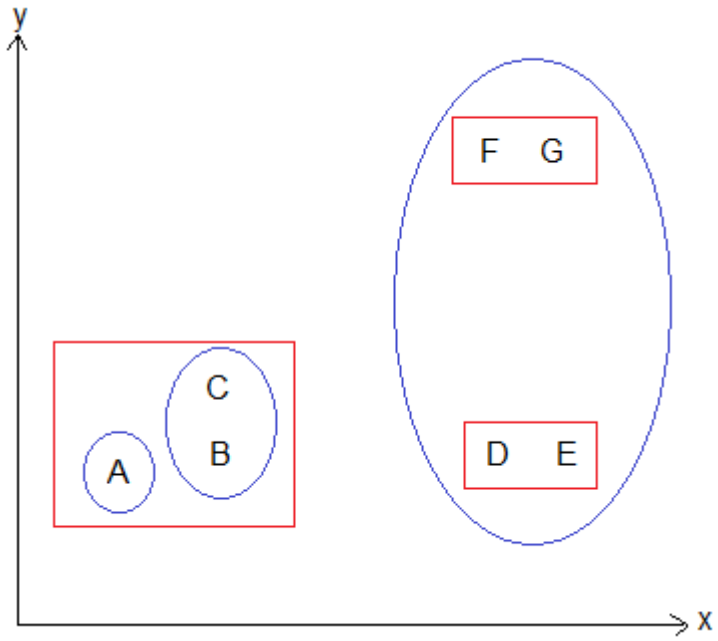
O algoritmo mais conhecido dessa categoria é o algoritmo K-means (também chamado de K-médias). Segundo Jain et. al, (1999, p.293) o algoritmo K-means “é

popular devido a sua facilidade de implementação e sua ordem de complexidade O(n), onde n é o número de padrões”. Devido a sua simplicidade na implementação, o algoritmo tornou-se o mais utilizado no contexto de problemas relacionados a agrupamento de dados, servindo de parâmetro para funcionamento de outros algoritmos de agrupamentos que surgiram posteriormente.

Seu procedimento segue uma maneira simples de classificar um determinado conjunto de dados por meio de um número de grupos (K) fixado a priori. Segundo Wives (2004, p.46) no K-means "o usuário indica o número de conglomerados desejado e o algoritmo de particionamento cria (de fora aleatória ou por outro processo) um conjunto inicial de partições (conglomerados)". No entanto, conforme Jain et.al.(1999, p.293) “um dos maiores problemas deste algoritmo é que o mesmo é sensível à seleção da partição inicial e pode convergir a um mínimo local do valor da função de critério se a o número de conglomerados inicial não for devidamente escolhida”. Um exemplo desse problema é apresentado a seguir:

> “Na Figura 13 mostra sete padrões bidimensionais. Se iniciarmos com os padrões A, B, e C como as médias iniciais em torno das quais os três clusters são construídos, então nós finalizamos com a partição {{A}, {B, C}, {D, E, F, G}} mostrada pelas elipses. O valor do critério de erro dentro dos retângulos é muito maior para esta partição que para a melhor partição {{A, B, C}, {D, E}, {F, G}} mostrada pelos retângulos, que engloba o valor global mínimo da função critério de erro que está nos retângulos para um agrupamento contendo três clusters. A solução de três clusters correta é obtida escolhendo, por exemplo, A, D, e F como as médias de cluster iniciais.” (Jain et. al., 1999 p. 276).

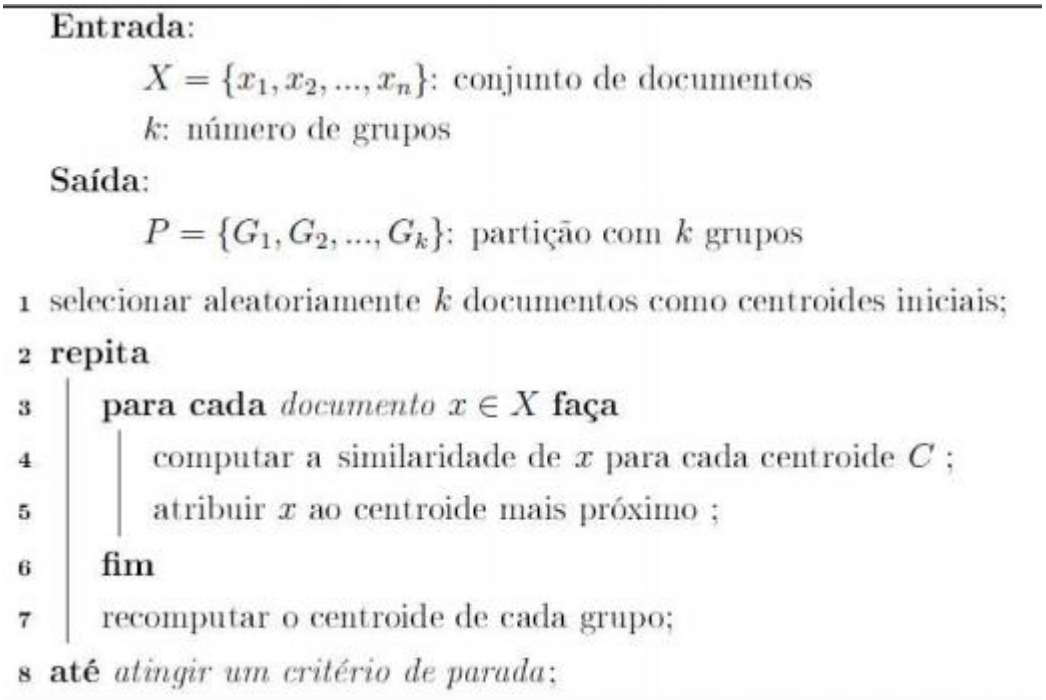
Figura 13. Sensibilidade do K-Means a Partição Inicial



Fonte: Jain et. al. 1999, p. 276

Definido os centroides, o K-Means associará todos os objetos aos clusters, comparando os objetos aos centroides que cada cluster possui. Os centroides representam uma categoria, ou seja, um grupo dos objetos dos clusters. A associação entre os objetos e os centroides é realizada por meio do calculo da "distância" entre os objetos e o centroide, e para isso, são usadas funções chamadas de funções de similaridade. Na função de similaridade pode ser utilizada a função co-seno, onde a distância é medida dividindo-se o produto de dois vetores pelo produto dos seus tamanhos. As etapas do algoritmo K-Means são apresentadas no pseudocódigo na Figura 14, abaixo.

Figura 14. Funcionamento do Algoritmo K-Means

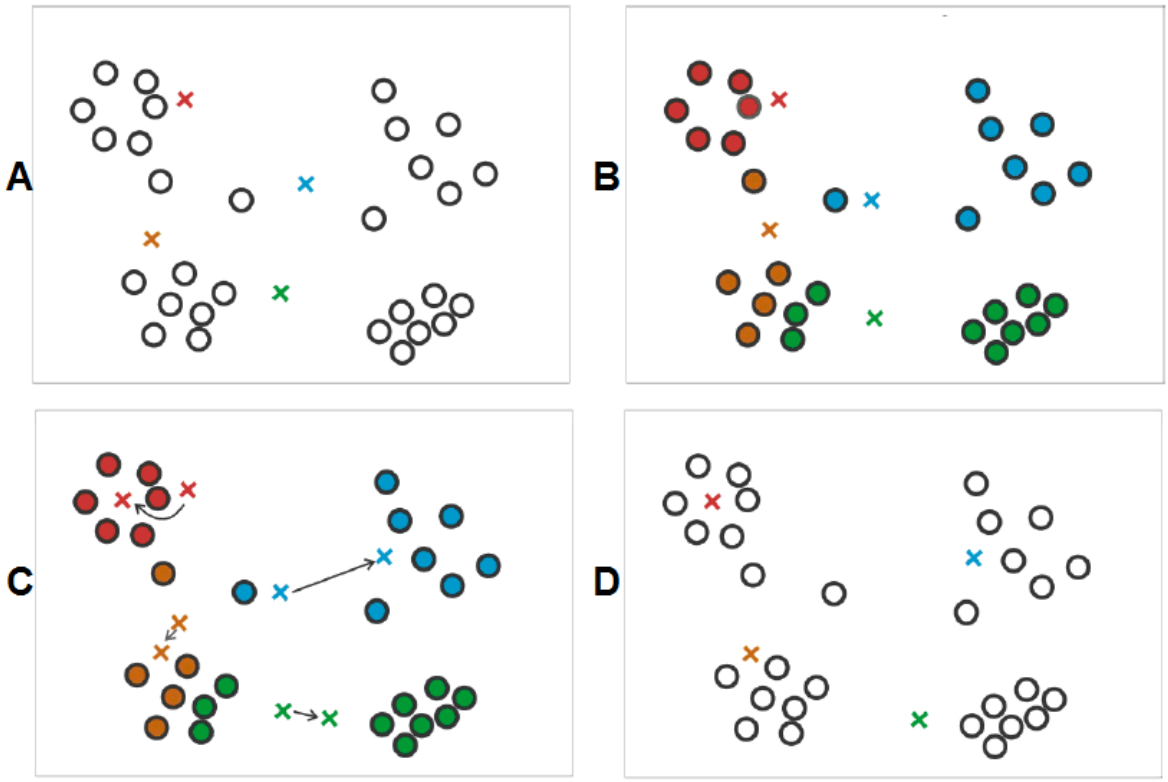


Conforme pseudocódigo na Figura 14, o algoritmo K-Means inicia seu processamento a partir da seleção dos centróides (linha 1), onde cada centróide representa um cluster. Cada cluster possui um centróide, para o qual deve ser aplicada as medidas de similaridades, onde compara o centróide com os demais itens do grupo. Conforme já mencionado, a escolha do número inicial de centróides é uma tarefa difícil, uma vez que dessa escolha, depende o sucesso do algoritmo. Uma das formas de escolher o número inicial de clusters é a forma randômica. Outra forma é fazer múltiplas iterações do algoritmo, com diferente quantidade de clusters, e então selecionar a iteração que produzir a maior coesão. A partir da definição do centróide, o algoritmo analisa a distância ou similaridade dos pontos selecionados com todos os itens a serem agrupados. Dessa forma, cada item é alocado ao cluster cujo centróide estiver mais próximo (linha 5). Um método para calcular a similaridade entre um item e o centróide consiste no cálculo do co-seno. A medida de similaridade co-seno varia entre 0 (nenhum termo comum) e 1 (todos os termos possuem o mesmo peso), que calcula a similaridade a partir do ângulo formado entre o vetor do cluster e o vetor do item comparado. O cluster cujo centróide for mais próximo ao item comparado, será o escolhido. Após cada atribuição de item (documento) a um cluster (associado a um centróide), esse centróide é recalculado para refletir esse novo cluster (atualizado) (linha 7). Em linhas gerais, o centróide é

um vetor que representa uma média dos pesos, de termo em termo de todos os itens que pertencem ao cluster. Esse processo continua até que nenhum centróide mude de posição.

O procedimento descrito acima pode ser visto em mais detalhes através da Figura 16 que mostra uma representação dos objetos em um plano cartesiano de duas dimensões.

Figura 15. Ilustração do Algoritmo K-Means



Para Stakoviak (2011, p. 36) "o fato dos métodos de particionamento fazerem várias passagens (iterações) no conjunto de dados, é considerado sua maior vantagem", obtendo melhores resultados. Isso se dá pelo fato da possibilidade de correção de eventuais alocações de um objeto a um grupo, ser inadequadas, sendo corrigida por essas iterações.

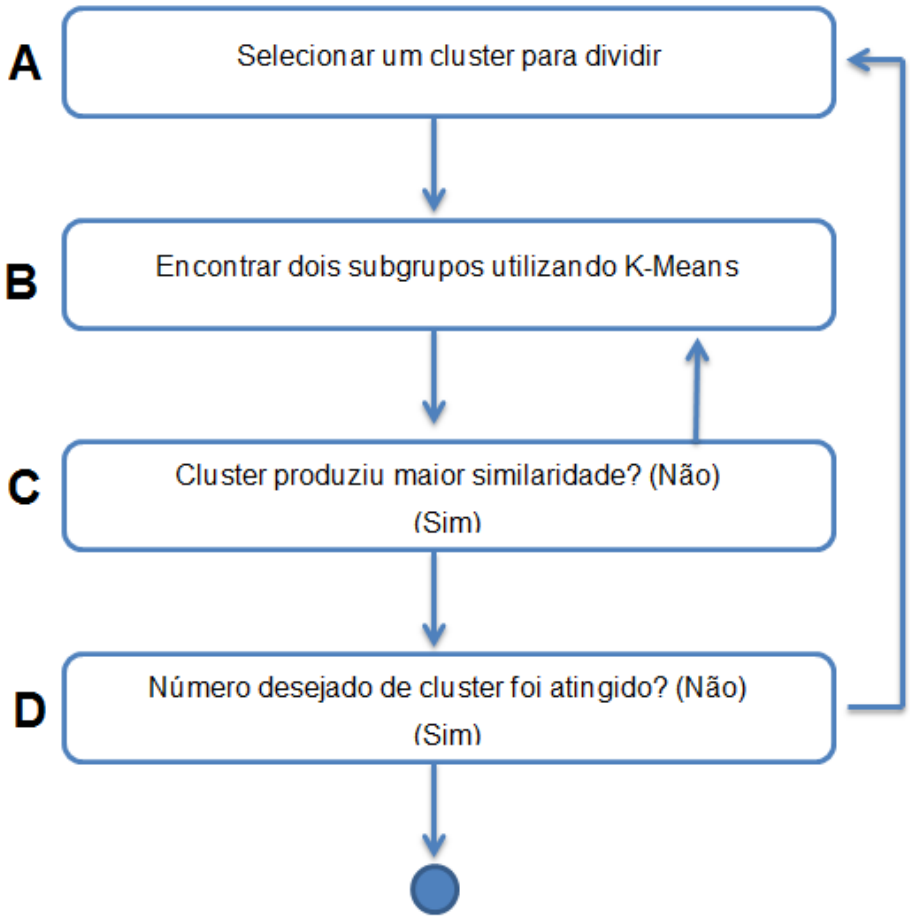
Na seção a seguir será apresentado o algoritmo *Bisecting* K-Means (ou K-Means Bisseccionado), uma das variações do K-Means.

2.4.2 K-Means Bisseccionado

Há diversas variações do algoritmo K-Means, utilizando diferentes estratégias e técnicas para agrupar elementos similares em classes. O algoritmo K-Means Bisseccionado consiste em uma variação hierárquica do algoritmo K-Means, que a cada iteração, seleciona um grupo e o divide, de forma a gerar uma hierarquia. (FONTANA & NALDI, 2009, p. 23). Os algoritmos hierárquicos criam uma hierarquia de relacionamentos entre os elementos, na qual, produzem um conjunto de dados agregados em que os clusters são unidos até que todo o conjunto de dados esteja ligado entre si.

Este algoritmo começa com um único cluster contendo todos os documentos e trabalha conforme apresentada na Figura 16.

Figura 16. Funcionamento do Algoritmo K-Means Biseccionado

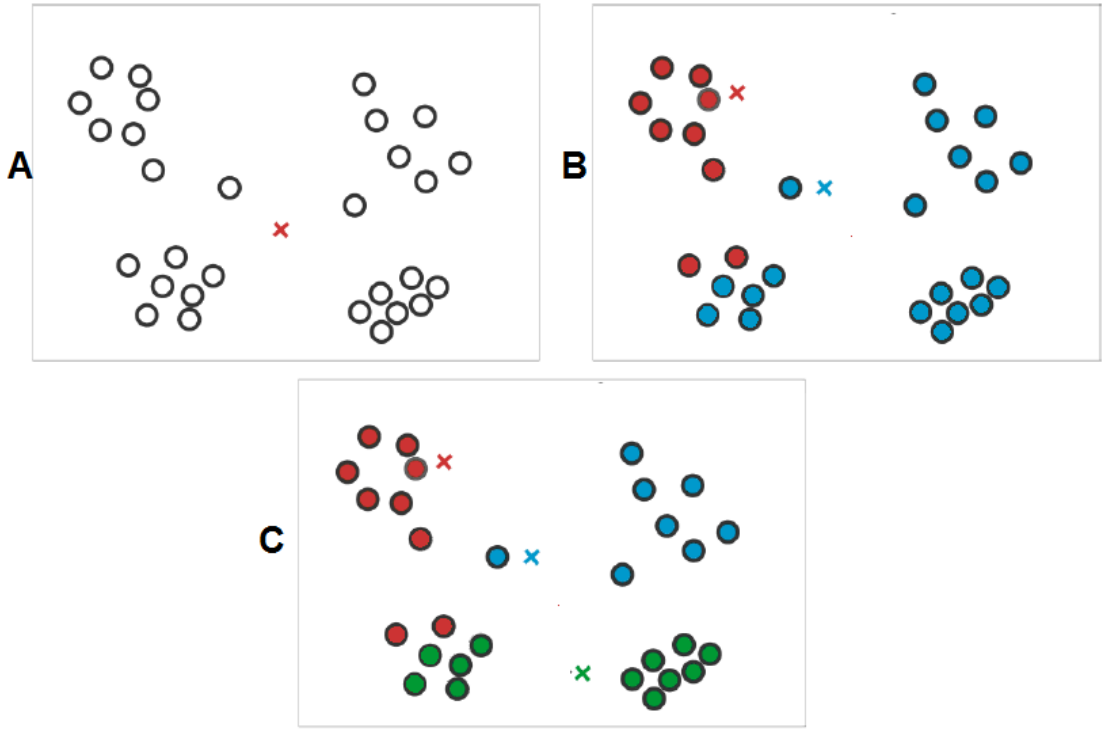


Conforme apresentado na Figura 16, o algoritmo O K-Means Biseccionado inicia seu processamento com um único cluster (Figura 17A) selecionado da divisão

em dois subgrupos do cluster inicial. O passo seguinte, consiste em dividir o cluster em dois novos subgrupos e executar o algoritmo tradicional do K-Means com parâmetro k igual a 2 para os dois subgrupos formados (Figura 17B). O algoritmo K-Means será executado até que se obtenha o cluster de elementos de maior similaridade (Figura 17C). Obtendo o cluster de maior similaridade, é verificada se a quantidade de agrupamentos desejada foi alcançada (Figura 17D). Sendo alcançada a quantidade de agrupamentos desejada, o algoritmo é encerado. Caso contrário, repetem-se os passos executados pelo algoritmo desde o início (Figura 17A).

O procedimento descrito acima pode ser visto em mais detalhes através da Figura 18 que ilustra as duas primeiras iterações do algoritmo K-Means Bisseccionado para uma base de dados.

Figura 17. Ilustração das Duas Primeiras Iterações do Algoritmo K-means Bisseccionado



Na Figura 17 (A) tem-se a inicialização da partição com um único cluster contendo todos os objetos. Em seguida (Figura 17B), o cluster inicial é dividido em dois novos clusters e aplicado o K-Means com parâmetro k igual a 2 em ambos os clusters. Seleciona-se um cluster (Figura 17B – cluster representado em azul) e divide-se em dois novos cluster aplicando o K-Means, também com parâmetro k igual a 2. Há diversas formas diferentes de selecionar qual cluster será dividido. Para Fontana & Naldi (2009, p. 23) a seleção do cluster dá-se de diversas formas,

“seja pelo maior tamanho (e.g., número de objeto ou diâmetro) ou menor similaridade entre os objetos e centróide (e.g. ou volume)”, no entanto, segundo Steinbach, Karypis & Kumar (2000, p. 11), "a diferença entre as formas de selecionar o cluster a ser dividido é pequena". Importante ressaltar que nas demais iterações do algoritmo (não ilustradas), os *clusters* continuarão a ser divididos até que o número de *clusters* desejado seja alcançado.

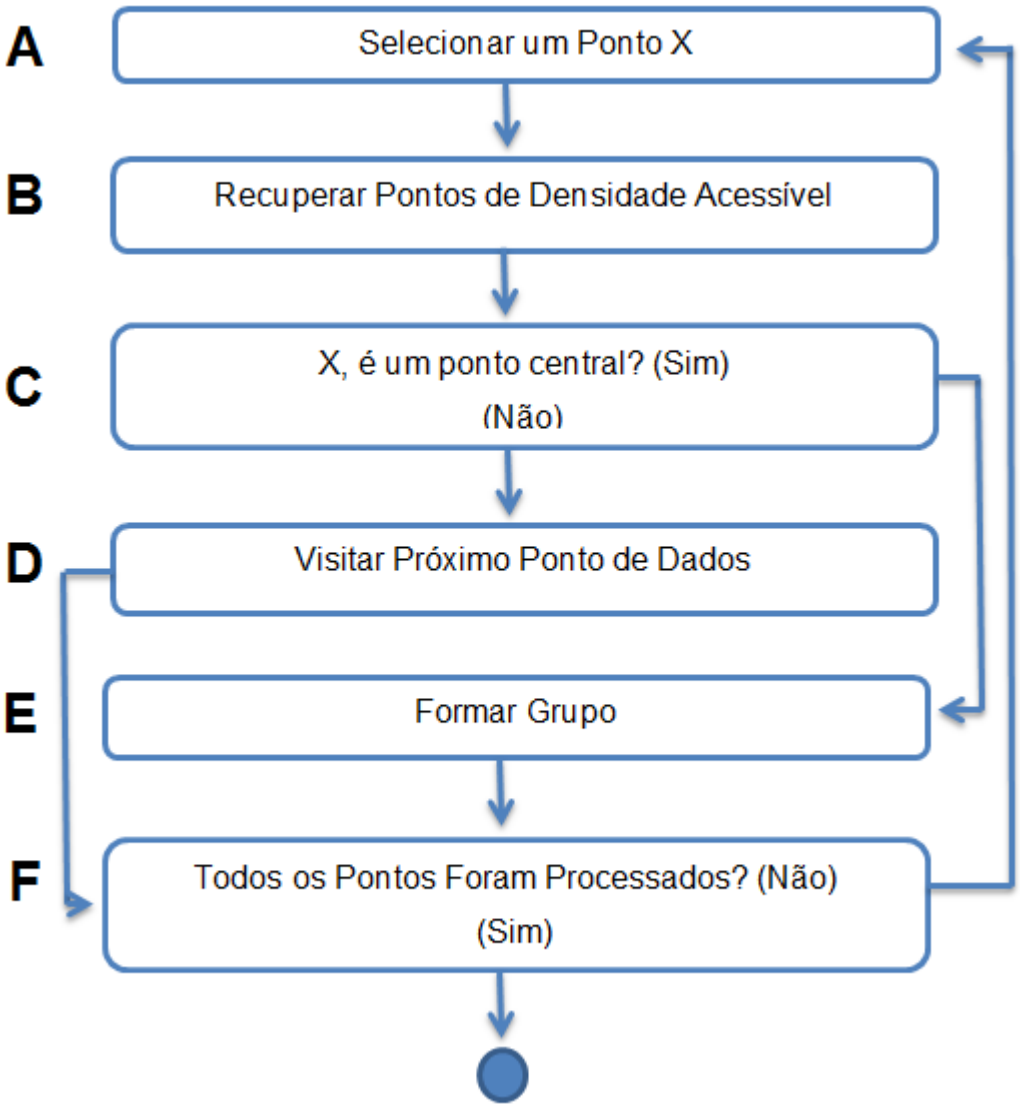
É possível observar que o K-Means Biseccionado apresenta uma complexidade computacional proporcional ao número de objetos, isso porque se o algoritmo for executado até formar clusters com um único objeto, o algoritmo realiza N-1 divisões durante sua total execução (uma divisão para cada objeto da base de dados a mais que um). Steinbach, Karypis & Kumar (2000, p. 13) "aponta que se o número de *clusters* é amplo e o refinamento não é utilizado, então k-means biseccionado é mais eficiente que o algoritmo k-means tradicional", isso porque o K-Means Biseccionado não possui a necessidade de comparar cada ponto a cada centróide, conforme execução do algoritmo K-Means tradicional, mas sim aos pontos do cluster e sua distância até os dois pontos centrais (centróides).

2.4.3 DBSCAN

O *Density-based Spatial Clustering of Applications with Noise* (DBSCAN) é um modelo de clusterização baseado na densidade, em que define clusters criando relações entre os objetos que se encontram nas áreas mais densas da região, isto é, forma grupos em limite mais irregulares. Nessa abordagem, é computada a alcançabilidade de um ponto a partir de um ponto inicial (semente), e então conecta os pontos alcançáveis com suas respectivas sementes. E conforme Tavares (2012 p. 27) "o DBSCAN é um dos algoritmos de *clustering* mais usados".

O algoritmo DBSCAN encontra clusters de objetos de forma arbitrária em bases de dados espaciais na presença de ruídos, sendo capaz de descobrir grupos de pontos de máxima densidade de ligações. "Os clusters são considerados regiões mais densas dos objetos espaciais de dados, sendo separados por regiões de baixa densidade ou ruído" (STAKOVIAK, 2011, p. 38). As etapas do algoritmo DBSCAN são apresentadas na Figura 18, abaixo.

Figura 18.Funcionamento do Algoritmo Density-Based



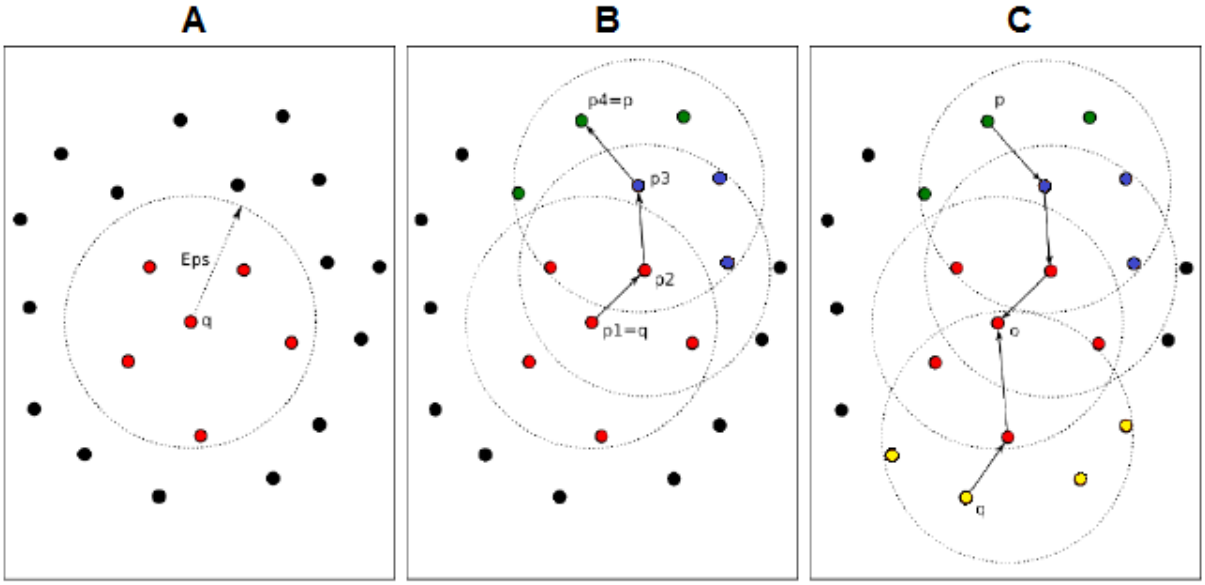
Fonte: STAKOVIAK, 2011, p. 38.

Conforme observado na Figura 18, seu processamento se inicia a partir da seleção arbitrária de um ponto X (Figura 18A). A etapa seguinte consiste na recuperação de todos os pontos de densidade acessível (Figura 18B) e então, é feita a verificação se o ponto X_1 é um ponto central, que possuem número mínimo de pontos para criar clusters ou mais vizinhos similares; ou de fronteira, que possuem pelo menos um vizinho classificado como ponto central (Figura 18C). Caso nenhum ponto seja acessível a partir da densidade (X_1 seja um ponto da fronteira), o DBSCAN visita o próximo ponto de dados do banco de dados (Figura 19D). Caso X_1 seja um ponto central, forma-se um grupo (Figura 18E). Um grupo é definido a partir de um conjunto de pontos que cumpram uma propriedade de conectividade, que

requer dois parâmetros, como conceito para formação de grupos: um valor máximo de distância entre pontos para que estes possam pertencer ao mesmo cluster (*Eps*) e um número mínimo de pontos necessários para se criar cluster (*MinPts*). Com esses parâmetros, dois pontos são alcançáveis se for possível ligá-los por meio de uma cadeia de pontos que cumpram esses parâmetros. Até que todos os pontos tenham sido processados (Figura 18F) o algoritmo reinicia seu processamento (Figura 18A).

Na Figura 19 é ilustrado o conceito de formação de grupos do algoritmo DBSCAN.

Figura 19. Conceitos Básicos do algoritmo DBSCAN



Um ponto é selecionado (Figura 19A) e o valor *Eps* delimita a distância máxima entre pontos, definindo um grupo. O próximo ponto é visitado e caso seja considerado ponto central, é formado o grupo, do contrário, um novo ponto é visitado (Figura 19B). Um grupo é definido por um conjunto de pontos que cumpram a propriedade de conectividade entre o *Eps* e *MinPts* (Figura 19C).

Uma das vantagens na utilização do DBSCAN é o fato de não ser necessário definir a priori, o número de clusters, com isso, é possível que o mecanismo se adapte a diferentes ambientes, bem como reduzindo a necessidade de interação em alterações no número de objetos disponíveis ao logo do tempo. Para Rehman (2006, p.2) outra característica que torna o DBSCAN apelativo como algoritmo de *clustering*

é a noção de ruído que permite remover objetos não desejados dos clusters ou agrupar todos os objetos considerados ruídos num cluster específico. Stakoviak (2011, p.39) ressalta que o DBSCAN, mesmo para grandes bases de dados espaciais é considerado um método eficaz no que tange a descoberta de cluster.

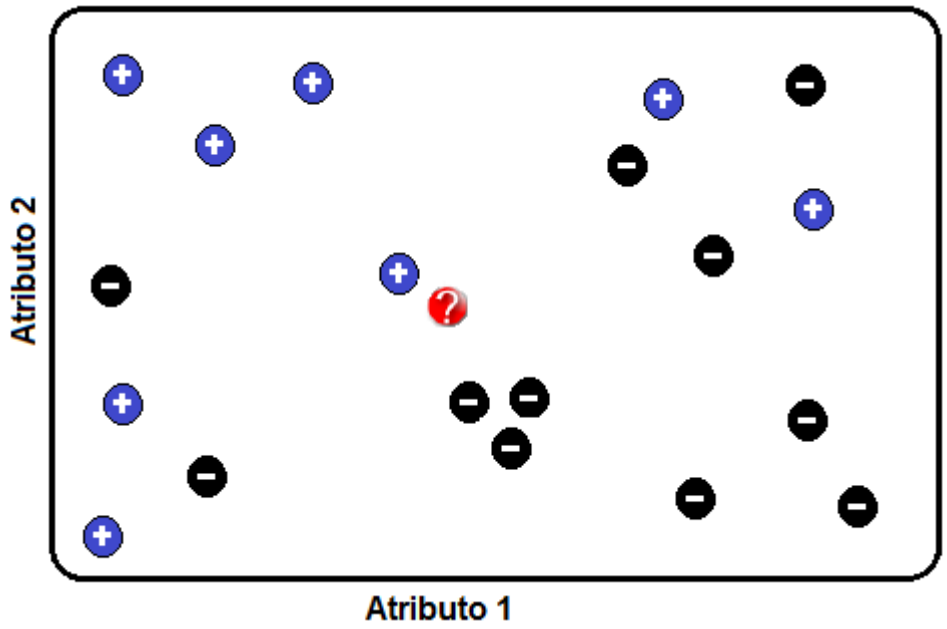
2.4.4 KNN

O algoritmo KNN (K *Nearest Neighbors*, K vizinhos Mais Próximos) é um algoritmo de aprendizagem supervisionado, baseado na analogia. No KNN, "uma maneira de predizer o valor y de um novo exemplo consiste em comparar esse exemplo com outros cuja classe é conhecida e atribuir à classe do caso mais próximo" (FERRERO, 2009, p.29). A esses exemplos conhecidos e o novo exemplo, compõem o conjunto de treinamento, sendo formado por, por vetores n-dimensionais e para cada elemento deste conjunto, é representado por um ponto no espaço n-dimensional.

O KNN é um classificador em que possui apenas um parâmetro, para definir o número de K-vizinhos, sendo o mesmo controlado pelo usuário com o intuito de obter uma melhor classificação. Este parâmetro é chamado K, o qual indica o número de vizinhos que serão usados pelo algoritmo durante a fase de teste. Para Bezerra (2006, p.31) "o parâmetro K faz com que algoritmo consiga uma classificação mais refinada, porém o valor ótimo de K varia de problema para o outro".

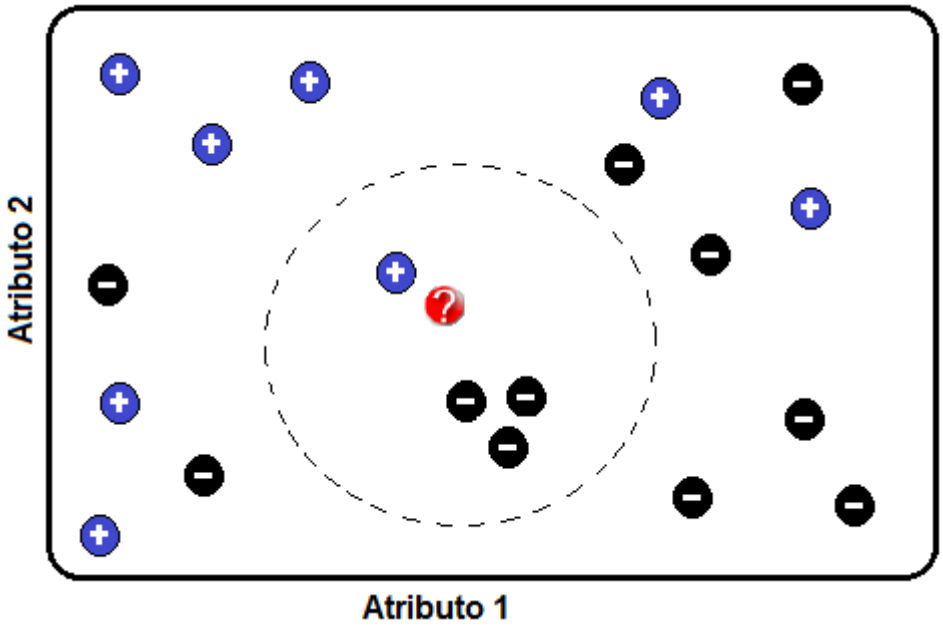
Segundo Ferrero (2009, p.29) "a ideia geral desse algoritmo consiste em encontrar os K exemplos rotulados mais próximos do exemplo não classificado e, com base no rótulo desses exemplos mais próximos, é tomada a decisão relativa a classe do exemplo não rotulado". Neste algoritmo, é determinado um volume V que contém os K-vizinhos mais próximos centrados em um padrão X, o qual se deseja classificar. Para aferir a classe de um novo padrão X, o algoritmo calcula os K-vizinhos mais próximos a X, ou seja, que tenha a menor distância, e classifica-o a classe com maior frequência dentre os seus K-vizinhos. A Figura 20 apresenta um exemplo de classificação deste algoritmo.

Figura 20. Exemplo de Classificação do Algoritmo KNN, com k=1



Na Figura 20 é apresentado um problema de classificação, com um conjunto de exemplos de treinamento descrito por dois atributos, onde atributos com rótulo positivos representa um atributo e exemplos com rótulo negativo representam outro atributo. Utilizando o KNN para classificação com K=1, o rótulo não classificado seria classificado de acordo com o único vizinho mais próximo, que é da classe positivo (+), baseado no cálculo de similaridade. No entanto, se k>1, então são considerados as classes dos k exemplos mais próximo para realizar a classificação. Na Figura 21, o k=4 e a maioria dos 4 exemplos mais próximos é negativo, e por isso, o rótulo não classificado, será classificado como negativo (-).

Figura 21. Exemplo de Classificação do Algoritmo KNN, com k=4



Conforme é possível observar, o número de vizinhos mais próximos a serem considerados na classificação de um novo rótulo, influencia fortemente a classificação. Não existe um único valor de K que seja apropriado para todos os problemas e cada problema em particular, deve ser avaliado.

No que tange a similaridade dos rótulos, o cálculo da menor distância é usualmente usado em um conjunto de dados descrito por atributos numéricos, assim a menor distância corresponde a maior similaridade. Diversos índices de proximidade têm sido propostos para o cálculo da similaridade entre dois pontos, sendo que a mais utilizada, é a distância Euclidiana. Abaixo se tem as métricas mais comuns no cálculo de distância (FERRERO, 2009, p.31).

Seja X=(x_1, x_2,...x_n) e Y=(y_1, y_2,..., y_n):

A distância Euclidiana entre X e Y é dada por:

$$\mathrm{d(x,y)} = \sqrt{(x_1 - y_1)^2 + (x_2 - \mathrm{y})^2 + \ldots (x_n - y_n)^2}$$

A distância de Manhattan entre X e Y é dada por:

$$\mathrm{d(x,y)} = |x_1 - y_1| + |x_2 - y_2| + \cdots |x_n - y_n|$$

A distância Minkowski entre X e Y é dada por

$$\mathrm{d(x,y)} = \ (|x_1 - y_1|^q + |x_2 - y_2|^q + \cdots |x_n - y_n|^q)^{\frac{1}{q}}$$

A distância de Minkowski é a generalização das distâncias Euclidiana e Manhattan. Quando q=1, esta distância representa a distância de Manhattan e quando q=2, a distância Euclidiana.

O processo de *clustering* pode ser computacionalmente exaustivo se considerar um conjunto com muitos dados. Para Ferrero (2009, p.29) "o custo computacional para rotular um novo exemplo é relativamente alto, pois, no pior dos casos, esse exemplo deverá ser comparado com todos os exemplos contidos no conjunto de exemplos de treinamento". Em contrapartida, o mesmo autor, relata que o algoritmo KNN querer pouco esforço durante a etapa de treinamento.

As técnicas de agrupamento de dados funcionam através da identificação de grupos de objetos que apresentam características semelhantes. No contexto deste trabalho, técnicas de *clustering* propõe identificar publicações semelhantes e após a identificação de grupos de publicações similares, a recomendação será direcionada mais fortemente.

3 METODOLOGIA

Para o desenvolvimento desse trabalho foram utilizados diversos recursos bibliográficos, *hardware* e *software*, que, aliados às orientações, permitiram a finalização do mesmo.

3.1 Materiais

Os materiais utilizados no desenvolvimento deste trabalho podem ser divididos em duas partes:

- Fontes bibliográficas: consiste nos materiais utilizados no desenvolvimento deste trabalho, como: dissertações, artigos, teses, monografias e publicações científicas;
- Softwares: para desenvolvimento do *plug-i*n foi utilizado o ambiente de desenvolvimento *NetBeans* IDE 8.1, juntamente com a API *WordPress Codex* e a linguagem de programação PHP, para implementação das etapas de *clustering* e recomendações.

3.1.1 APIs WordPress

Application Programming Interface (API), ou Interface de Programação de Aplicativos, em português, são métodos desenvolvidos que formam um conjunto de funções que são disponibilizados para que outros softwares possam utilizá-las, onde o software que utiliza essa API não precisa envolver-se com detalhes da implementação da mesma, mas apenas utilizar suas funcionalidades.

O *WordPress* é uma ferramenta para publicação e gerenciamento de conteúdo na web que mais têm crescido nos últimos anos, proveniente da sua arquitetura altamente extensível através de *plug-ins*. Um *plug-in* no *WordPress* são códigos escrito em PHP que agrega um conjunto específico de recursos ou serviços para o blog do *WordPress* e que é incluído na pasta *wp-content/plugins,* aumentando *as* funcionalidades do *WordPress*. Uma das vantagens em se utilizar *plug-ins* no *WordPress*, consiste na flexibilidade de funcionalidades em que é possível implementar, além das rotinas implementadas não sofrerem com futuras atualizações da versão do *WordPress*.

O *WordPress* fornece várias APIs para que os *plug-ins* possam trabalhar usando-as. Cada API interage de diferentes maneiras. A seguir é descrito algumas APIs:

- API *Plug-in* - providencia um conjunto de ações e filtros para que os plug-ins possam interagir em várias partes do *WordPress*.
- API *Database* - permite acesso rápido a base de dados do *WordPress* sem que haja necessidade de conhecimento em SQL, permitindo inserir, atualizar, remover e selecionar registros.
- API Settings - responsável pela inserção de campos de opções personalizados na administração do *WordPress*, com o intuito de usar os mesmo facilmente pela *plug-in*.
- API *Dashbord Widgets* - API responsável por criar e alterar *widgets* para a *Dashboard* da administração, bem como o controle do acesso dos *widgets*.

3.2 Metodologia

Diversas pesquisas foram realizadas, de maneira que permitissem oferecer uma sustentação teórica necessária para o desenvolvimento do presente trabalho. Dessa forma, foram abordados conceitos e técnicas que fosse possível desenvolver um mecanismo cujo objetivo, possibilitasse a recomendação de publicações em um *plug-in* no *WordPress*, como forma de concretização dos conceitos explanados na Revisão de Literatura.

O projeto está dividido em duas etapas. A etapa inicial consiste no estudo teórico dos conceitos envolvidos, definição dos objetivos e justificativos do projeto e a construção de um referencial teórico abordando estes conceitos. Os conceitos estudados na etapa inicial foram:

- Recuperação da Informação (RI): seção responsável por apresentar o conceito de RI e SRI. No entendimento dessa seção é possível compreender o processo de recuperação da informação, bem como é feito esse processo e sua importância para o sistema de recomendação.
- Sistemas de Recomendação (SR): nesta seção é apresentando uma abordagem geral a certa dos conceitos de um SR e as técnicas clássicas de recomendação, tais como FBC, FC e FH.

- Técnicas de Recomendação: essa seção apresenta uma explanação na técnica de recomendação baseada em conteúdo, uma vez que se trata da técnica utilizada no trabalho. Essa etapa também envolveu o estudo dos algoritmos de *clustering* apresentados neste trabalho. No entendimento dessas técnicas é possível compreender suas utilizações e finalidades

Ainda nesta fase, partiu-se para a compreensão mais detalhada das APIs do *WordPress*, realizando alguns testes em um servidor local, pois só a partir da realização desses testes seria possível compreender detalhadamente o funcionamento das rotinas de implementação de *plug-ins* para *WordPress*. No estudo sobre Técnicas de Recomendação Baseado em Conteúdo, foi necessário entender conceitos sobre Recuperação da Informação, como os modelos clássicos e a etapa de preparação dos documentos, como *stopwords* e *stemming*. Além disso, foram realizados testes em algumas das técnicas abordadas na seção 2.4, como o K-*Means* e o *K-Means Bisseccionado*. Portanto, essa etapa foi primordial para o andamento do trabalho, pois são as técnicas que realizam o processo de identificação da similaridade das publicações, sendo assim o ponto principal do Sistema de Recomendação proposto e o ponto chave do trabalho.

Ainda nesta etapa, foram definidos os objetivos e justificadas do trabalho. Esta compreensão foi muito importante para definir os limites deste trabalho, assim como o que será entregue como resultado final.

Por fim, finalizando a primeira etapa, foi elaborada a revisão de literatura. O conhecimento adquirido a partir dos estudos realizados serviu como base para a definição e realização das próximas etapas do projeto.

A segunda etapa do projeto consiste no desenvolvimento do projeto. Nesta etapa envolveu a escolha das linguagens de programação, IDE (*Integrated Development Environment*) e SGBD utilizado. Além disso, foi definido o processo de desenvolvimento da Filtragem Baseado em Conteúdo, utilizando conceitos de clustering para agrupamentos das publicações mais similares e posteriormente a Distância Euclidiana para gerar as publicações ordenadas mais similaridades.

Para que o projeto fosse desenvolvido, foram necessários os passos:

- Definir o modelo de dados da aplicação responsável por gerar as recomendações baseadas em conteúdo;
- Definir uma técnica de clusterização, dentre as abordagens estudadas;
- Implementar o algoritmo de clustering;
- Agregar o algoritmo de *clustering* ao *plug-in* e realizar testes para validar os resultados obtidos.

4 RESULTADOS E DISCUSSÃO

Os conceitos estudados tiveram como objetivo obter a fundamentação teórica para permitir que o mecanismo de recomendação desenvolvido pudesse recomendar publicações utilizando técnica de Filtragem Baseada em Conteúdo.

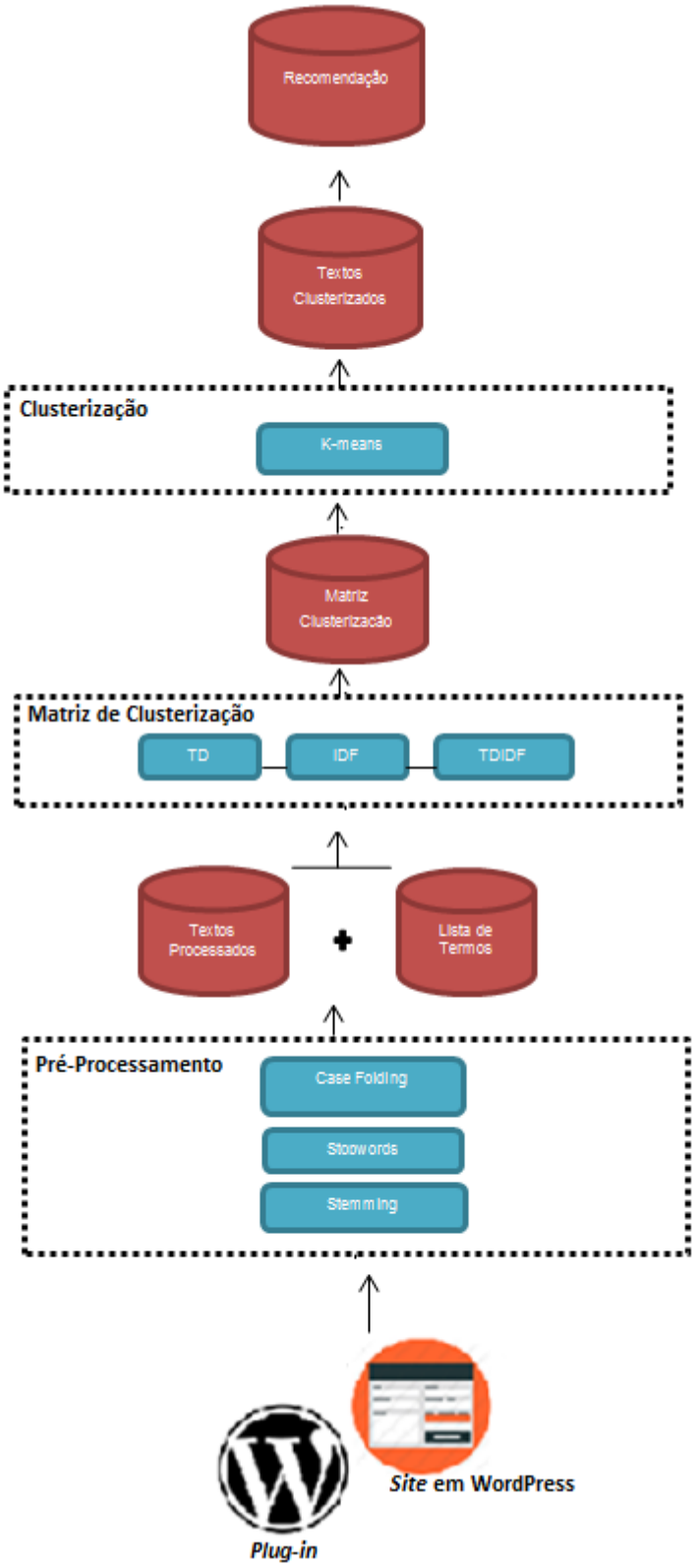
Este capítulo é voltado para a apresentação do processo de desenvolvimento do *plug-in* responsável por apresentar as recomendações aos usuários, bem como do sistema de recomendação, responsável por identificar e gerar as recomendações baseada na similaridade textual dos marcadores existentes nas publicações.

Nas próximas seções serão apresentadas o fluxo de funcionamento do Mecanismo de Recomendação, a aplicação cliente, o modelo do usuário, implementação do sistema de recomendação e testes.

4.1 Fluxo de Funcionamento do Mecanismo de Recomendação

O mecanismo desenvolvido gera as recomendações analisando apenas as informações contidas nas publicações existentes no *site* desenvolvido na Plataforma WordPress. Dessa forma, o *plug-in* proposto poderá ser acoplado a qualquer *site* em *WordPress*, servindo como módulo adicional para geração das recomendações. A Figura 22, a seguir, apresenta o fluxo de funcionamento do mecanismo de recomendação.

Figura 22. Fluxo Básico de Funcionamento do Mecanismo de Recomendação



A estrutura do mecanismo de recomendação ilustrado na Figura 22 funciona da seguinte forma: Inicialmente o *plug-in* é instalado no *site* desenvolvido em

WordPress. Durante a instalação, são criadas as tabelas que fornecem suporte para o funcionamento do Sistema de Recomendação (seção 4.2).

Na tela de administração do *plug-in* o administrador do *site* seleciona a quantidade de agrupamentos gerados (número de *clusters*), a quantidade de recomendações que serão apresentadas e o período durante o qual o Sistema de Recomendação irá gerar as similaridades entre as publicações. Isso é feito para que as recomendações sejam realizadas de forma *off-line* e não necessite gerar as recomendações a cada visita em uma publicação no *site*. O modelo de recomendação *off-line* consiste em gerar as recomendações periodicamente e armazená-las para que no momento em que houver a necessidade de apresenta-las, o mecanismo apenas consume esses registros e apresenta ao usuário. Além disso, as informações salvas na base de dados terão uma validade conforme informada nas configurações, passado este prazo, o mecanismo irá gerar uma nova geração de agrupamento e atualizar a tabela na base de dados.

Em seguida, o *plug-in,* utilizando a API do *WordPress*, recupera as informações das publicações (título e *tags*) e as salva nas tabelas criadas (base de dados do serviço de recomendação). O Sistema de Recomendação utiliza as informações das publicações para aplicar os conceitos de Recuperação da Informação (etapa de Pré-Processamento) citados na seção 2.1.2. Após isso, é gerada uma matriz que dá suporte ao processo de *clustering*. Esta matriz é criada utilizando o TF-IDF tendo como entrada as *tags* existentes nas publicações. Uma vez produzida a matriz de *clustering*, passa-se ao processo de *clustering*, utilizando o algoritmo K-Means. Ao final do processo de *clustering*, é gerada uma lista contendo os agrupamentos formados com as publicações. Esses agrupamentos servem de base para refinamento do processo de recomendação, uma vez que traz os itens mais similares. Após isso, é criada uma lista de publicações mais similares utilizando a fórmula do co-seno e essa lista é armazenada em uma tabela na base de dados. Sempre que um usuário visualizar uma publicação, o *plug-in* irá buscar na base de dados as publicações, ordenando as mais similares e apresentará como recomendação.

A próxima seção apresenta a forma como a base de dados da aplicação foi organizada.

4.2 Base de Dados da Aplicação

Com o objetivo de otimizar o funcionamento do *plug-in*, são criadas 4 tabelas na aplicação, no momento em que o *plug-in* é instalado. Estas tabelas são criadas utilizando a API do *WordPress*, responsáveis pela manipulação e persistência dos dados. A base de dados da aplicação armazena informações sobre o *site*, bem como suas publicações e assim, gera as recomendações e as salva em uma tabela para que o *plug-in* busque as publicações similares e recomenda ao usuário no momento em que está acessando uma publicação. É a partir desse modelo que as recomendações são geradas de forma *off-line*. A Figura 23 apresenta a estrutura do modelo lógico do mecanismo.

Figura 23. Modelo Lógico do Mecanismo

A base de dados é constituída pelas seguintes tabelas:

- PluginClustering_Administracao - contém as configurações das recomendações no *website*, tais como a quantidade de categorias (agrupamentos), a quantidade de recomendações a ser apresentado, o período para gerar as recomendações e a data de instalação do *plug-in*.
- PluginClustering_Recomendacao – contém um id, o id das publicações salvos na base de dados do *WordPress*, o número do agrupamento atribuído a publicação e o fator de similaridade. As publicações similares serão compostas do mesmo valor do campo “Agrupamento” e o fator de similaridade é usado para retornar as publicações mais similares.

Conforme pode ser observado na Figura 23, a base de dados da aplicação não armazena informações sobre as publicações, uma vez que, utilizando as API do *WordPress* é possível recuperar todas as informações referentes as publicações, como título, marcadores e outras.

A estrutura de base de dados descrita nesta seção oferece ao mecanismo a possibilidade de armazenar as recomendações e servir de suporte para que a aplicação cliente (*plug-in*) busque as recomendações e apresente ao usuário. A próxima seção descreve sobre a aplicação cliente, responsável pela apresentação das recomendações aos usuários.

4.3 Aplicação Cliente

A aplicação cliente consiste no *plug-in* desenvolvido, utilizando as APIs do *WordPress*. Após a instalação do *plug-in*, o administrador do *site* deve configurar o número de categorias a ser criado (número de *cluster*), número de recomendações a ser apresentado ao usuário e o período com o qual o Sistema de Recomendação deve analisar a similaridade das publicações. Para isso, o administrador do *site* deve se direcionar a página de administração dentro do *menu* "Configurações" e adequar as configurações a realidade do site.

Figura 24. Tela de administração do plug-in

Conforme apresentado na Figura 24, após a instalação do *plug-in,* o administrador do *site* deve realizar as configurações adequadas à realidade da

aplicação. No entanto, essas configurações não é um critério obrigatório, uma vez que o *plug-in* está configurado por padrão a criar 8 (oito) categorias, apresentar 4 (quatro) publicações como o número de recomendações geradas e “uma semana” como o período para processar as recomendações. Apenas o número de *cluster* é um critério fortemente recomendado para configuração, sendo relativo às necessidades de cada aplicação.

Para implementação da Tela de Administração do *plug-in* foi utilizada a API do *WordPress*.

Figura 25. Código-fonte Tela Inicial do Plug-in

```
27    //Tela de Administração
28    add_action('admin_menu', 'gera_menus');
29    function gera_menus(){
30        add_options_page('Recomendação de Posts', 'Recomendação FBC', 'administrator', 'recomendacao', 'fbc_config_adm');
31    }
32    function fbc_config_adm(){
33        if (isset($_POST['save_config']) && $_POST['save_config']=='confirm'):
34            update_option('rec_fbc_numeroRecomendacaoGeradas', $_POST['fbc_numeroRecomendacaoGeradas']);
35            update_option('rec_fbc_periodoRecomendacao', $_POST['fbc_periodoRecomendacao']);
36            $aviso = '<div class="updated"><p><strong>Suas configurações foram salvas com sucesso!</strong></p></div>';
37        endif;
38        ?>
39        <div class="wrap">
40            <?php if(isset($aviso)) echo $aviso; ?>
41            <h2>Configuração Global do Serviço de Recomendação Baseado em Conteúdo</h2>
42            <form method="post" action="<?php echo $_SERVER['REQUEST_URI'] ?>">
43                <h3>Configure seus anúncios abaixo</h3>
44                <table class="form-table">
45                    <tr valign="top">
55                    <tr valign="top">
66                    </tr>
67                </table>
68                <p class="submit">
69                    <input type="submit" value="Salvar Configurações" class="button-primary" name="salvar" />
70                    <input type="hidden" value="confirm" name="save_config" />
71                </p>
72            </form>
73        </div>
74        <?php
75    }
```

A função "fbc_config_adm" (linha 32) implementa toda a rotina de apresentação da tela de administração fazendo uso de HTML e funções do PHP . A tela de administração é apresentada usando a função "gera_menus" (linha 29), que insere a página de opções no menu de Configurações. Na função "gera_menus" é usado a função do *WordPress* "add_options_page" (linha 30) reponsável por adicionar uma pagina na estrutura do *WordPress*, passando como parâmetro o título da página, o nome do que irá aparecer no menu, o nível de usuário para obter acesso a página, o slug da página e a função que irá gerar o menu.

Feita as configurações na Tela de Administração, o *plug-in* está configurado e pronto para uso. Assim, o *plug-in* irá gerar a similaridade entre as publicações conforme o período definido e irá salvar as similaridades em uma tabela específica

no banco de dados da aplicação. No momento em que um usuário acessar uma publicação, o *plug-in* irá buscar a lista de publicações mais similares a publicação alvo da recomendação e irá apresentar ao usuário.

Com a implementação do *plug-in* finalizada, o próximo passo a ser realizado consistiu na implementação do Sistema de Recomendação, que será abordado na próxima seção.

4.4 Sistema de Recomendação

O Sistema de Recomendação desenvolvido tem o objetivo de recomendar publicações que de acordo com a similaridade entre as *tags* (marcadores) descritas em cada publicação. O sistema de recomendação proposto neste trabalho utiliza técnica de *clustering* para agrupar as publicações mais similares e assim servir de base para o processo de recomendação, utilizando técnicas de Filtragem Baseada em Conteúdo. Portanto, para iniciar o processo de recomendação, o *plug-in* irá recuperar todas as publicações existentes no *site e*, assim, dar-se-á início ao processo de *clustering*.

Clustering é uma técnica bastante utilizada para agrupar documentos similares. O processo inicial consiste na criação dos vetores de características referentes a cada publicação, neste caso, foram utilizados os marcadores das publicações. O próximo passo é determinar a similaridade das publicações baseado em seus vetores de características, utilizando o TF-IDF descrito na seção 2.1.2. Para este cálculo, o conjunto de marcadores (itens) é representado por um vetor de frequência destes itens e os pesos de um vetor de características definem um ponto no espaço e localizam a menor distância entre esses pontos retornando os itens mais similares. Uma vez que a distância entre as características que descrevem as publicações é calculada, eles podem ser agrupados utilizando a técnica de *clustering* mais apropriada.

O algoritmo utilizado neste trabalho foi o K-means, por ser um dos mais tradicionais algoritmos de *clustering*. A Figura 27 apresenta parte do código da classe K-means desenvolvida em PHP.

Figura 26. Implementação da Classe K-means

```
1  <?php
2  namespace webd\clustering;
3  class KMeans
4  {
5      public $k = 4;
6      public $n = 9;
7      public $pontos = array();
8      public $centroide = array();
9
10     public function executar()
11     {
12         $this->EscolhaInicialCentroides();
13         for ($i = 0; $i < $this->n; $i++)
14         {
15             $this->removerPontos();
16             $this->AtribuirPontos();
17             $this->ComputarNovosCentroides();
18             if ($this->TestarSeConvergiu())
19             {
20                 break;
21             }
22         }
23     }
```

No k-means, o usuário indica o número de *clusters* desejado e o algoritmo cria de forma aleatória um conjunto inicial de partições. Conforme pode ser observado na Figura 27, na classe desenvolvida o número mínimo de clusters é fixado em 4, o número máximo de iterações é fixado em 9 (linha 6), são criados os *arrays* dos pontos e os centroides, e os centroides iniciais são escolhidos aleatoriamente (linha 12). A seguir, o centroide de cada um desses *clusters* é computado e o algoritmo analisa a distância desses com todos os elementos a serem agrupados (linha 16). Após, cada elemento é alocado ao *cluster* cujo centroide esteja mais próximo e, ao ser incluído, este centroide é re-computado (linha 17) para representar esse novo elemento. O processo é repetido até que os centroides não mudem mais de posição (linha 18).

O algoritmo K-means é inicializado pela associação randômica de itens aos clusters. Os itens selecionados são definidos como os centroides iniciais. No K-means, o centróide de um cluster é definido como o vetor médio de todos os itens num cluster para cada dimensão separadamente. A Figura 28 apresenta a função que gera os valores iniciais dos centroides, escolhendo os k itens randomicamente e associando-se cada um deles a um cluster diferente.

Figura 27. Função de Inicialização dos Centróides

```php
protected function EscolhaInicialCentroides()
{
    for ($i=0; $i<$this->k; $i++)
    {
        $this->centroide[] = $this->pontos[mt_rand(0, count($this->pontos)-1)]->converterParaCentroide();
    }
}
```

Após a escolha aleatória da posição dos centroides, a etapa seguinte consiste em cada interação, os valores das posições dos demais itens seja removida para que em seguida, possa ser recalculada a distância dos pontos a dos centroides. Em seguida, para cada conjunto de *tags* (item) que descrevem um *post* é calculado a distância Euclidiana entre a posição do item e cada centróide. A distância Euclidiana tem por objetivo obter um valor que determina a similaridade entre dois objetos. A similaridade entre um item e um centróide é calculada como o somatório de todos os vetores item no cluster dividido pelo número de vetores. A Figura 29 apresenta a função de atribuição das posições.

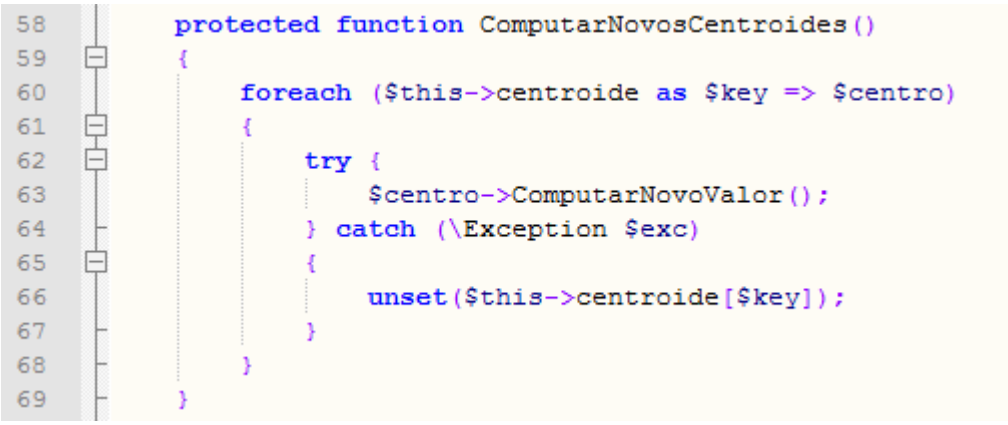
Figura 28. Função de Atribuição das Posições

```php
protected function AtribuirPontos()
{
    foreach ($this->pontos as $ponto)
    {
        $mais_curta = 0;
        $distance_mais_curta = PHP_INT_MAX;
        foreach ($this->centroide as $centroide_id => $centro)
        {
            $distancia = $ponto->distanciaPara($centro);
            if ($distancia < $distance_mais_curta)
            {
                $distance_mais_curta = $distancia;
                $mais_curta = $centroide_id;
            }
        }

        $this->centroide[$mais_curta]->addPonto($ponto);
    }
}
```

Quando todos os itens forem devidamente realocados entre os clusters, calcula-se os novos centroides para cada um dos clusters. A Figura 30 apresenta a função para geração de novos centroides.

Figura 29. Função para Geração de Novos Centróides

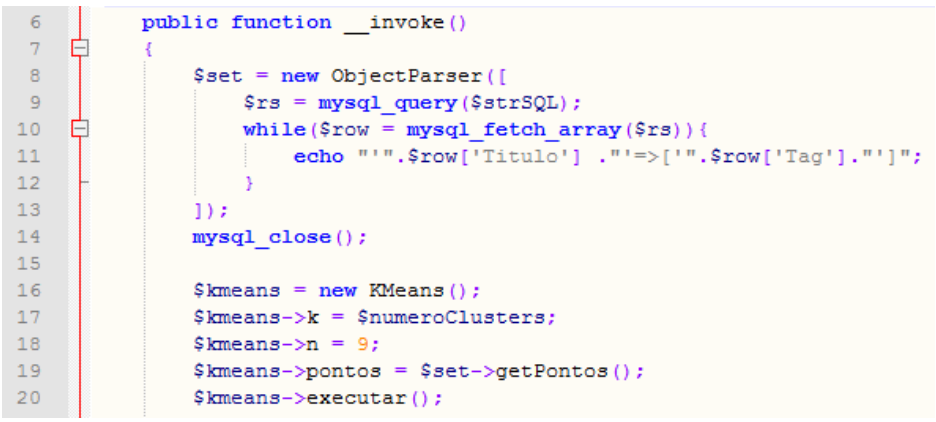


```
58      protected function ComputarNovosCentroides()
59      {
60          foreach ($this->centroide as $key => $centro)
61          {
62              try {
63                  $centro->ComputarNovoValor();
64              } catch (\Exception $exc)
65              {
66                  unset($this->centroide[$key]);
67              }
68          }
69      }
```

Conforme pode ser observado na Figura 30, a cada centroide é chamado à função responsável por computar o novo valor para a posição (linha 63). O processo continua até atingir o número definido de interações ou até quando não houver mais mudanças nas posições.

Concluído a implementação do algoritmo K-means, a Figura 31 apresenta a função responsável pela inicialização do *clustering*.

Figura 30. Inicialização da Clustering utilizando o algoritmo K-means



```
6       public function __invoke()
7       {
8           $set = new ObjectParser([
9               $rs = mysql_query($strSQL);
10              while($row = mysql_fetch_array($rs)){
11                  echo "'".$row['Titulo'] ."'=>['".$row['Tag']."']";
12              }
13          ]);
14          mysql_close();
15
16          $kmeans = new KMeans();
17          $kmeans->k = $numeroClusters;
18          $kmeans->n = 9;
19          $kmeans->pontos = $set->getPontos();
20          $kmeans->executar();
```

Conforme pode ser observado na Figura 31, nas linhas 10 e 11 é realizada uma busca na base de dados e retornado o atributo “*Tags*” dos *posts*. O valor da partição inicial é selecionado pelo valor atribuído ao administrador do *website* nas configurações do *plug-in* e o número de interações ficou definido como 9 (linhas 17 e 18). De posse dessas informações é possível iniciar o algoritmo e gerar os agrupamento, isso é feito nas linhas 19 a 20. Note que no algoritmo deve-se definir previamente o número k de clusters, pois não se sabe quantos conjuntos serão precisos para definir as regiões de interesse dos usuários. A escolha de um valor

muito alto ou muito baixo para o parâmetro k pode levar o sistema a tomar decisões equivocadas, atribuindo um item a um cluster que com estas configurações estaria melhor representando, porém no contexto geral poderia ser atribuído a outro cluster. Foi selecionado 9 iterações por considerar um número significativo, já que é necessário calcular diversas vezes a função de distância até que haja estabilização dos valores dos centroides.

Com os *cluster* formado e de posse dos valores de similaridade entre os itens existentes em cada *cluster*, uma lista de recomendação é gerada utilizando a fórmula do co-seno e armazenada essas recomendações em uma tabela no banco de dados. Dessa forma, no momento em que um usuário acessar uma publicação, o *plug-in* irá buscar a lista de publicações mais similares a publicação alvo da recomendação e irá apresentar ao usuário, evitando que ocorra um novo processamento para geração das recomendações. Vale ressaltar que essas recomendações serão válidas pelo tempo informado no menu de administração do *site*. Passado o prazo informado, será feita uma nova geração de agrupamento e atualizado a tabela na base de dados. Na próxima seção é apresentado um teste realizado em ambiente local sobre o *plug-in* de recomendação desenvolvido.

4.5 Teste

Nesta seção são apresentados os resultados da aplicação *clustering K-means* como mecanismo para calcular a similaridade entre descritores presente nas publicações na Plataforma *WordPress*. Este algoritmo solicita ao usuário o número de *clusters* desejados e organiza os documentos (publicações na Plataforma *WordPress*) de forma a serem alocados no número de *clusters* desejado com base na distância entre os termos que descrevem os vetores das publicações.

Para este experimento foi utilizado o *site* (En)Cena[1] que consiste em um portal para o qual convergem produções textuais referentes ao tema da loucura e possui um acervo de mais de 1.200 trabalhos de pesquisadores, acadêmicos, profissionais e usuário do sistema de saúde, artistas entre outros ((En)Cena, 2016, online). O (En)Cena foi escolhido como ambiente de teste por apresentar um

[1] http://encenasaudemental.net/series/

conjunto de séries, estruturado em formas de seções e cada artigo possuir um conjunto de marcadores (*tags*) que descrevem cada uma das publicações.

Os passos para a realização do experimento foram os seguintes:

1. Extração do título e *tags* das publicações no site (En)cena. Foram extraídos somente o título e os marcadores das publicações porque estes normalmente descrevem de forma bem objetiva e abrangente as publicações. Vale observar que todas as publicações possuíam *tags*.
2. Instalação da Plataforma *WordPress* em um ambiente local para simulação e inserção dos dados coletados na etapa anterior no ambiente criado. Para o ambiente de teste foi utilizado o Xampp e a inserção dos dados coletados foi realizada de forma manual.
3. Execução do experimento.

O ambiente de simulação foi montado a partir do site (En)Cena contendo diferentes seções e ao total foram extraídos 223 publicações. O *site* disponibiliza 22 seções (séries) diferentes cada uma contendo um conjunto de publicações, ao qual cada publicação possui um conjunto de *tags* que descrevem a publicação. Os assuntos disponíveis, bem como a quantidade de publicações para as respectivas seções e a quantidade geral de marcadores existentes para cada assunto são mostrados a seguir.

Tabela 2. Características do corpus (En)Cena.

Seção (série)	**Número de Publicações**	**Número Total de *Tags***	**Número Médio de Tags por Publicação**
As Bruxas dos Contos de Fadas	8	41	5,13
Budismo e Cristianismo	9	35	3,89
Contemporaneidade Líquida	7	28	4
Demônio em Fuga	6	28	4,67
Deuses Gregos	19	128	6,74

Encena em Campo	7	35	5
Fragmentos do Saber	7	36	5,14
Harry Potter – O processo de individuação do herói	7	32	4,57
Humano Demasiado Tecnológico	7	27	3,86
La Fora	6	26	4,33
Mitologia Africana	16	80	5
Mulheres Modernas	7	35	5
Mundo do Trabalho e Sofrimento Psiquco	7	34	4,86
Oscar 2013	11	43	3,91
Oscar 2014	11	49	4,45
Oscar 2015	20	94	4,7
Oscar 2016	17	111	6,53
Poder Subjetividade Saber	12	73	6,08
Princesas Disney	14	67	4,79
Saúde Mental Indígena	9	52	5,78
Sete Pecados Capitais	8	34	4,25
Sete Virtudes	8	31	3,88
Total	**223**	**1119**	
Média	**10,14**	**50,86**	**5,02**

A Tabela 3 apresenta os códigos das publicações presentes em cada uma das seções. Essa estrutura foi elaborada tendo-se como base a tabela de seções com as

características do ambiente teste (Tabela 2). Os códigos das publicações contêm uma letra S no sentido de “Seção”, um número ordenado conforme distribuição alfabética das seções, uma letra P fazendo referência a “Publicação” e um número inserido de forma ordenada conforme listagem anterior, sem qualquer dependência ao site (EN)Cena. O título e os marcadores das publicações podem ser obtidos na página do site, mas também pode ser encontrado no Anexo, para facilitar a análise e a avaliação dos resultados.

Tabela 3. Distribuição das Publicações por Seção no site (EN)Cena

Seção (Série)	Publicações
As Bruxas dos Contos de Fadas	S1P1, S1P2, S1P3, S1P4, S1P5, S1P6, S1P7, S1P8
Budismo e Cristianismo	S2P9, S2P10, S2P11, S2P12, S2P13, S2P14, S2P15, S2P16, S2P17,
Contemporaneidade Líquida	S3P18, S3P19, S3P20, S3P21, S3P22, S3P23, S3P24
Demônio em Fuga	S4P25, S4P26, S4P27, S4P28, S4P29, S4P30
Deuses Gregos	S5P31, S5P32, S5P33, S5P34, S5P35, S5P36, S5P37, S5P38, S5P39, S5P40, S5P41, S5P42, S5P43, S5P44, S5P45, S5P46, S5P47, S5P48, S5P49
Encena em Campo	S6P50, S6P51, S6P52, S6P53, S6P54, S6P55, S6P56
Fragmentos do Saber	S7P57, S7P58, S7P59, S7P60, S7P61, S7P62, S7P63
Harry Potter – O processo de individuação do herói	S8P64, S8P65, S8P66, S8P67, S8P68, S8P69, S8P70
Humano Demasiado Tecnológico	S9P71, S9P72, S9P73, S9P74, S9P75, S9P76, S9P77
La Fora	S10P78, S10P79, S10P80, S10P81, S10P82, S10P83
Mitologia Africana	S11P84, S11P85, S11P86, S11P87, S11P88, S11P89, S11P90, S11P91, S11P92, S11P93, S11P94, S11P95, S11P96, S11P97, S11P98,

	S11P99
Mulheres Modernas	S12P100, S12P101, S12P102, S12P103, S12P104, S12P105, S12P106
Mundo do Trabalho e Sofrimento Psiquco	S13P107, S13P108, S13P109, S13P110, S13P111, S13P112, S13P113
Oscar 2013	S14P114, S14P115, S14P116, S14P117, S14P118, S14P119, S14P120, S14P121, S14P122, S14P123, S14P124
Oscar 2014	S15P125, S15P126, S15P127, S15P128, S15P129, S15P130, S15P131, S15P132, S15P133, S15P134, S15P135
Oscar 2015	S16P136, S16P137, S16P138, S16P139, S16P140, S16P141, S16P142, S16P143, S16P144, S16P145, S16P146, S16P147, S16P148, S16P149, S16P150, S16P151, S16P152, S16P153, S16P154, S16P155
Oscar 2016	S17P156, S17P157, S17P158, S17P159, S17P160, S17P161, S17P162, S17P163, S17P164, S17P165, S17P166, S17P167, S17P168, S17P169, S17P170, S17P171, S17P172, S18P173, S18P174, S18P175
Poder Subjetividade Saber	S18P176, S18P177, S18P178, S18P179, S18P180, S18P181, S18P182, S18P183, S18P184
Princesas Disney	S19P185, S19P186, S19P187, S19P188, S19P189, S19P190, S19P191, S19P192, S19P193, S19P194, S19P195, S19P196, S19P197, S19P198
Saúde Mental Indígena	S20P199, S20P200, S20P201, S20P202, S20P203, S20P204, S20P205, S20P206, S20P207
Sete Pecados Capitais	S21P208, S21P208, S21P210, S21P211, S21P212, S21P213, S21P214, S21P215
Sete Virtudes	S22P216, S22P217, S22P218, S22P219, S22P220, S22P221, S22P222, S22P223

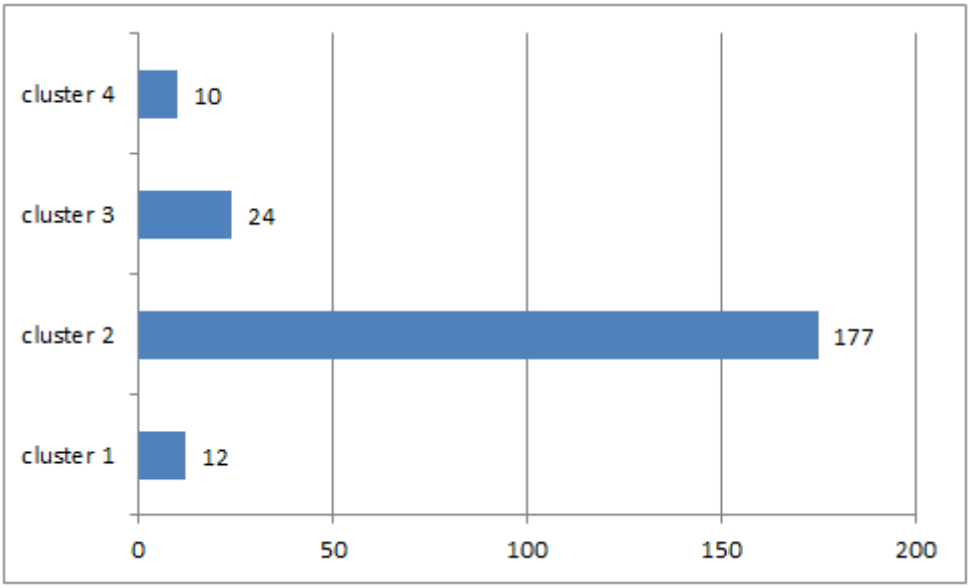
Ao todo, três experimentos foram realizados, a fim de avaliar a distribuição e a quantidade de documentos em cada *cluster*. O algoritmo K-means solicita ao usuário o número de *clusters* desejado e organiza os documentos de forma a serem alocados no número de *clusters* desejado com base na similaridade. Em todos os experimentos foram definidas 10 iterações no algoritmo e a geração de 4, 8 e 22 *clusters*

Os experimentos e suas peculiaridades são detalhados a seguir. Após o detalhamento, é apresentado o resumo dos resultados de todos os três experimentos.

4.5.1 Primeiro experimento

A Figura 32 a seguir mostra o resultados obtido para o conjunto de dados extraídos do site (En)Cena e exposto ao processo de *clustering* para geração de 4 *clusters*.

Figura 31. Distribuição Gerada com 4 Clusters



Conforme é possível observar na Figura 32, todos os *clusters* obtiveram um número considerado de publicações, podendo diversificar possíveis recomendações. No entanto, o resultado apresentou um *cluster* que obteve um número considerado de publicações similares comparados aos demais. A estrutura de clusters resultante para o experimento com 4 clusters é apresentado na Tabela 4.

Tabela 4. Estrutura de Cluster do Primeiro Experimento

Cluster	Publicações
Cluster 1	S18P173, S18P174, S18P175, S18P176, S18P177, S18P178, S18P179, S18P180, S18P181, S18P182, S18P183, S18P184
Cluster 2	S1P1, S1P2, S1P3, S1P4, S1P5, S1P6, S1P7, S1P8, S2P9, S2P10, S2P11, S2P12, S2P13, S2P14, S2P15, S2P16, S2P17, S4P25, S4P26, S4P27, S4P28, S4P29, S4P30, S5P31, S6P50, S6P51, S6P52, S6P53, S6P54, S6P55, S6P56, S7P57, S7P58, S7P59, S7P60, S7P61, S7P62, S7P63, S8P64, S8P65, S8P66, S8P67, S8P68, S8P69, S8P70, S9P71, S9P72, S9P73, S9P74, S9P75, S9P76, S9P77, S10P78, S10P79, S10P80, S10P81, S10P82, S10P83, S11P84, S11P85, S11P86, S11P87, S11P88, S11P89, S11P90, S11P91, S11P94, S11P99, S12P101, S12P103, S12P105, S12P106, S13P107, S13P108, S13P109, S13P110, S13P111, S13P112, S13P113, S14P114, S14P115, S14P116, S14P117, S14P118, S14P119, S14P120, S14P121, S14P122, S14P123, S14P124, S15P125, S15P126, S15P127, S15P128, S15P129, S15P130, S15P131, S15P132, S15P133, S15P134, S15P135, S16P136, S16P137, S16P138, S16P139, S16P140, S16P141, S16P142, S16P143, S16P144, S16P145, S16P146, S16P147, S16P148, S16P149, S16P150, S16P151, S16P152, S16P153, S16P154, S16P155, S17P156, S17P157, S17P158, S17P159, S17P160, S17P161, S17P162, S17P163, S17P164, S17P165, S17P166, S17P167, S17P168, S17P169, S17P170, S17P171, S17P172, S19P185, S19P186, S19P187, S19P188, S19P189, S19P190, S19P191, S19P192, S19P193, S19P194, S19P195, S19P196, S19P197, S19P198, S20P199, S20P200, S20P201, S20P202, S20P203, S20P204, S20P205, S20P206, S20P207, S21P208, S21P208, S21P210, S21P211, S21P212, S21P213, S21P214, S21P215, S22P216, S22P217, S22P218, S22P219, S22P220, S22P221, S22P222, S22P223
Cluster 3	S5P32, S5P33, S5P34, S5P35, S5P36, S5P37, S5P38, S5P39, S5P40, S5P41, S5P42, S5P43, S5P44, S5P45, S5P46, S5P47,

	S5P48, S5P49, S11P92, S11P93, S11P95, S11P96, S11P97, S11P98.
Cluster 4	S3P18, S3P19, S3P20, S3P21, S3P22, S3P23, S3P24, S12P100, S12P102, S12P104

Analisando o resultado obtido é possível constatar que os agrupamentos foram realizados de forma muito semelhante às atribuições das publicações nas séries no *site* (EN)Cena. O cluster 1 teve como centroide a publicação “S18P173” e agrupou somente elementos pertencentes a mesma série (“Poder Subjetividade Saber”). Os *clusters* 3 e 4 agruparam publicações pertencentes a duas séries cada, e o cluster 2 agrupou todas as demais publicações.

A Tabela 5 apresenta os termos mais frequentes no primeiro experimento. Essa medida é utilizada para identificar as palavras ou características mais importantes do cluster, que pode ser utilizado para compreendê-lo ou descrevê-lo.

Tabela 5. Termos mais Frequentes do Primeiro Experimento

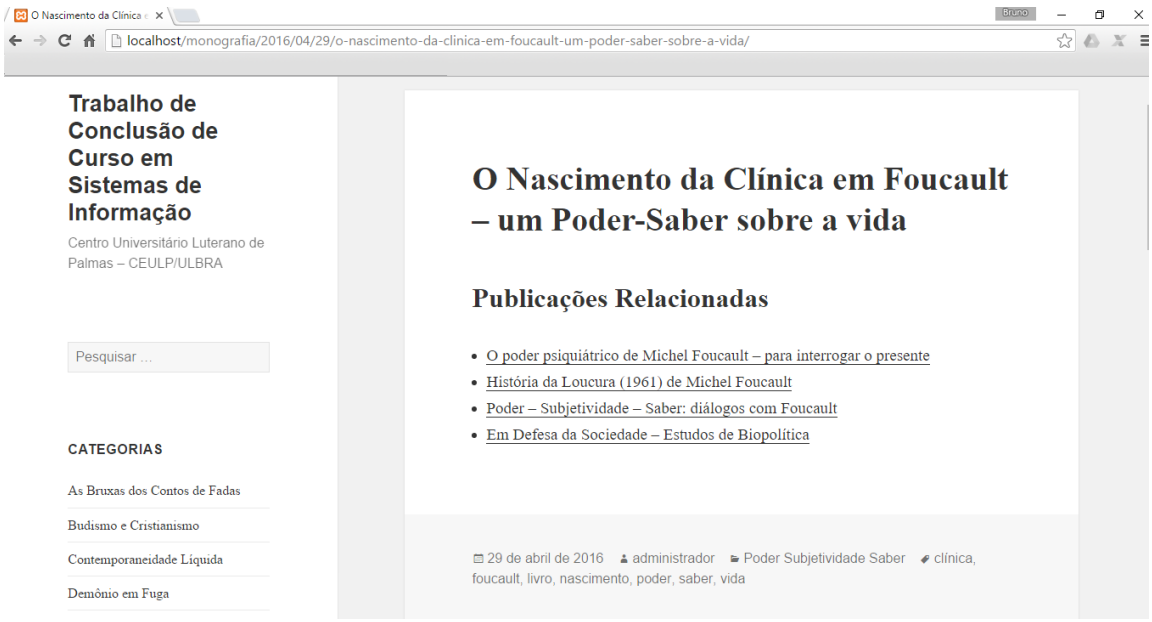
Cluster	**Termos mais frequentes (frequência de documentos)**
Cluster 1	foucault (12), livro (11), poder (4), sociedade (3), psiquiatria (2), sujeito (2), saber (2), história (2).
Cluster 2	cinema(56), religião(26), cultura(17), Disney(15), princesas(15), Oscar2015(15), arquétipo(14), Oscar2016(13), filosofia(12), estereótipos(12), Oscar2013(11), feminino(10), amor(10), pecado(10), Oscar2014(10), culpa(9), virtudes(8), budismo(8), cristianismo(8), fantasia(8), herói(7), orixá(7), extensão(6), conto de fadas(6), depressão(6), sofrimento(6), divindade(5), karajá(5), mulher(5), suicídio(4), Deus(4), vida(4), movimentos(4), moderna(4), trabalho(4), sexo(4), iny(3), fotografia(3), javaé(3), mito(3), EUA(3), bruxas(3), bruxa(3), mãe(3), inveja(3), discurso(3), conhecimento(3), demônio(3), luta(3), anatomia(3), morte(3), brasil(3), futebol(3), idolo(3), filósofo(3), liberdade(3), CEULP(3), comunidade(3), indígena(3), universidade(3), comportamento(3), ambíguo(3), ciclos(3) psique(3), vaidade(3), guerra(3), animus(2), alma(2),

	feminina(2), homem(2), existência(2), iluminismo(2), terapia(2), bipolar(2), psicopatologia(2), copa(2), esporte(2), ciência(2), harry(2), inconsciente(2), castidade(2), relacionamento(2), amizade(2), CIA(2), escravidão(2), bauman(2), animação(2), mental(2), saúde(2), psicodinâmica(2), poder(2), tragédia(2), Oscar(2), arquétipos(2), relação(2), luto(2), viagem(2), crescimento(2), resiliência(2), transformação(2), criança(2), infância(2), gênero(2), ação(2), xambioá(2), indígenas(2), jovens(2)
Cluster 3	analitica(16), mitologia(16), psicologia(15), deuses(11), grecia(7), Deus(5), grega(4), arquétipo(2), grego(2) gregos(2), jung(2)
Cluster 4	modernidade(10), bauman(7), liquida(7), cartas(2), crônica(2), mulher(2)

A frequência apresentada na Tabela 5 foi identificada com base na quantidade de termos em que apareceu no conglomerado e é apresentada entre parênteses ao lado de cada termo. Somente foram selecionados os termos cuja frequência apareceu mais de uma vez no conglomerado.

A Figura 32 apresenta o resultado da recomendação para o primeiro experimento.

Figura 32. Recomendação - Primeiro Experimento



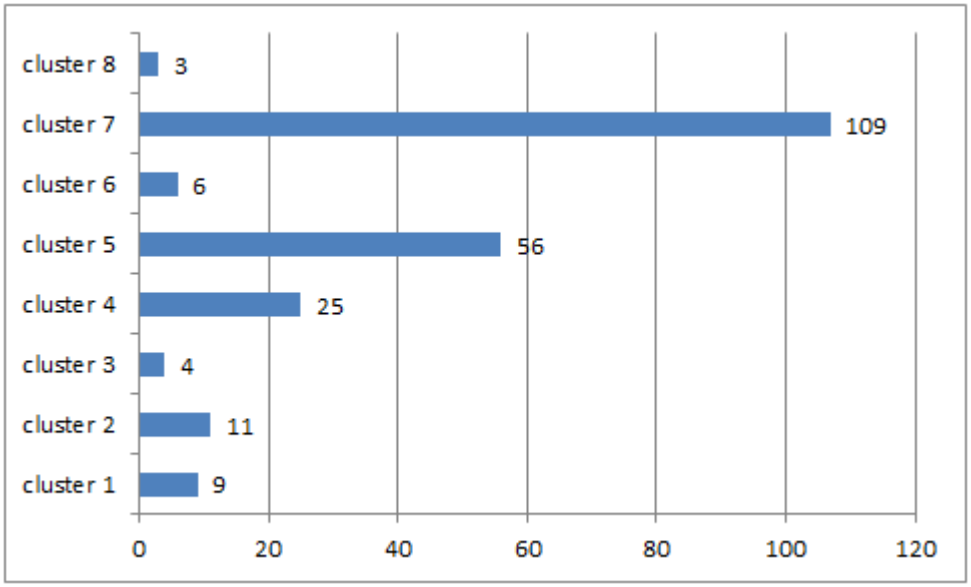
Conforme apresentado na Figura 32, após visita a publicação "O Nascimento da Clínica em Foucault – um Poder-Saber sobre a vida" é recomendada quatro publicações. As quatro publicações recomendadas fazem parte do mesmo cluster, conforme pode ser observado na Tabela 4.

A próxima seção apresenta os resultados para o segundo experimento.

4.5.2 Segundo experimento

A Figura 33 apresenta o resultado do segundo experimento, onde a quantidade de *cluster* foi definida como 8.

Figura 33. Distribuição Gerada com 8 Clusters



Analisando a Figura 33 é possível constatar que novamente houve um *cluster* que se destacou em quantidade de publicações similares, embora tenha agrupado menos termos similares, comparado ao experimento anterior. Essa diminuição pode ser explicada com o aumento do número de *clusters*, resultando em outros agrupamentos mais similares. Outro fator importante diz respeito aos *clusters* que obtiveram poucos elementos agrupados similarmente (*cluster* 3 e *cluster* 8). Para um mecanismo de recomendação que apresenta, por exemplo, quatro publicações como recomendação para o usuário, o *cluster* 8 não iria retornar todas as quatros publicações considerando apenas a técnica de *clustering*, necessitando de outra técnica para completar os itens da publicação.

A estrutura de *clusters* resultante para o experimento com 8 clusters é apresentado na Tabela 6.

Tabela 6. Estrutura de Cluster do Segundo Experimento

Cluster	Publicações
Cluster 1	S18P174, S18P175, S18P176, S18P177, S18P178, S18P180, S18P181, S18P183, S18P184
Cluster 2	S3P18, S3P19, S3P20, S3P21, S3P22, S3P23, S3P24, S12P100, S12P102, S12P104, S15P125
Cluster 3	S20P200, S20P201, S20P204, S20P205
Cluster 4	S2P9, S2P10, S2P11, S2P12, S2P13, S2P14, S2P15, S2P16,

	S7P58, S21P208, S21P208, S21P210, S21P211, S21P212, S21P213, S21P214, S21P215, S22P216, S22P217, S22P218, S22P219, S22P220, S22P221, S22P222, S22P223
Cluster 5	S8P64, S8P65, S8P66, S14P114, S14P115, S14P116, S14P117, S14P118, S14P119, S14P120, S14P121, S14P122, S14P123, S14P124, S15P126, S15P127, S15P128, S15P129, S15P130, S15P131, S15P132, S15P133, S15P134, S15P135, S16P136, S16P138, S16P139, S16P140, S16P141, S16P142, S16P143, S16P145, S16P146, S16P148, S16P149, S16P150, S16P151, S16P152, S16P153, S16P154, S16P155, S17P156, S17P157, S17P158, S17P159, S17P160, S17P161, S17P162, S17P163, S17P164, S17P165, S17P166, S17P167, S17P168, S17P170, S17P172,
Cluster 6	S4P25, S4P26, S4P27, S4P28, S4P29, S4P30
Cluster 7	S1P1, S1P2, S1P3, S1P4, S1P5, S1P6, S1P7, S1P8, S2P17, S5P31, S5P32, S5P33, S5P34, S5P35, S5P36, S5P37, S5P38, S5P39, S5P40, S5P41, S5P42, S5P43, S5P44, S5P45, S5P46, S5P47, S5P48, S5P49, S6P50, S6P51, S6P52, S6P53, S6P54, S6P55, S6P56, S7P57, S7P59, S7P60, S7P61, S7P62, S7P63, S8P67, S8P68, S8P69, S8P70, S9P71, S9P72, S9P73, S9P74, S9P75, S9P76, S9P77, S10P78, S10P79, S10P80, S10P81, S10P82, S10P83, S11P84, S11P85, S11P86, S11P87, S11P88, S11P89, S11P90, S11P91, S11P92, S11P93, S11P94, S11P95, S11P96, S11P97, S11P98, S11P99, S12P101, S12P103, S12P105, S12P106, S13P107, S13P108, S13P109, S13P110, S13P111, S13P112, S13P113, S16P137, S16P144, S16P147, S17P169, S17P171, S19P185, S19P186, S19P187, S19P188, S19P189, S19P190, S19P191, S19P192, S19P193, S19P194, S19P195, S19P196, S19P197, S19P198, S20P199, S20P202, S20P203, S20P206, S20P207
Cluster 8	S18P173, S18P179, S18P182

Coincidentemente, o centroide do cluster 1 foi novamente uma publicação da série “Poder Subjetividade Saber” e todas as publicações agrupadas neste *cluster*

pertencem a esta série. Este resultado pode ser explicado pela forte relação existente na similaridade dos marcadores existentes nas publicações desta série. De forma semelhante, os clusters 3 e 8 agruparam somente publicações pertencentes a mesma série. Em geral, é possível observar que os *clusters* são agrupados por publicações pertencentes à mesma distribuição das séries no site (EN)Cena.

Como no experimento anterior, os termos mais frequentes dos conglomerados foram identificados e apresentados na Tabela 7.

Tabela 7. Termos mais Frequentes do Segundo Experimento

Cluster	**Termos mais frequentes (frequência de documentos)**
Cluster 1	foucault(9), livro(8), poder(4), história(2), psiquiatria(2), sujeito(2), saber(2)
Cluster 2	modernidade(10), bauman(8), liquida(7), cartas(2), amor(2), crônica(2), mulher(2)
Cluster 3	cultura(4), fotografias(3), javaé(3), karajá(3), sofrimento(3), ação(2), xambioá(2)
Cluster 4	cultura(11), pecado(10), culpa(9), virtudes(8), budismo(8) cristianismo(8), filosofia(6), Deus(4), vaidade(3), discurso(3) homem(2), sexo(2), inveja(2)
Cluster 5	cinema(54), Oscar2015(15), Oscar2013(11), Oscar2016(10), Oscar2014(10), herói(5), amor(4), arquétipo(3), psique(3), guerra(3), EUA(3), morte(2), harry(2), porter(2), sexo(2), poder(2), CIA(2), tragédia(2), Oscar(2), tecnologia(2), crescimento(2), resiliência(2), transformação(2), criança(2) relação(2), luto(2), viagem(2), escravidão(2), animação(2)
Cluster 6	depressão(6), demônio(3), vida(3), anatomia(3), luta(2)
Cluster 7	mitologia(25), psicologia(22), analitica(18), arquétipo(17), Disney(14), princesas(14), deuses(13), estereótipos(12), orixá(12), feminino(9), fantasia(8), grecia(8), extensão(6), filosofia(6), tecnologia(6), jung(6), africana(5), divindade(5), conto de fadas(5), mãe(5), Deus(5), grega(5), amor(5), moderna(4), trabalho(4), mulher(4), harry potter(4), universidade(3), CEULP(3), comunidade(3), ambíguo(3),

	ciclos(3), liberdade(3), idolo(3), movimentos(3), deusa(3), herói(3), futebol(3), bruxas(3), sofrimento(3), suicídio(3), humano(3), bruxa(3), futuro(3), Oscar2016(2), iny(2), karajá(2), jovens(2), mental(2), mito(2), saúde(2), psicodinâmica(2), cultura(2), indígena(2), animus(2) feminina(2), grego(2), gregos(2), evolução(2), conhecimento(2), filósofo(2), brasil(2), copa(2), esporte(2)
Cluster 8	foucault(3), livro(3), sociedade(3)

Observando-se os dados apresentados na Tabela 7, pode-se perceber que no cluster 7 possui termos de mesmo significado, mas com diferentes terminações não sendo considerados idênticos, embora esteja no mesmo conglomerado. Os termos "grego", "gregos" e "grega" são um exemplo. Esta ocorrência faz com que as publicações que os possuem não sejam considerados similares como deveriam, em conjunto com outros termos, essas publicações poderiam ser alocadas em cluster diferentes.

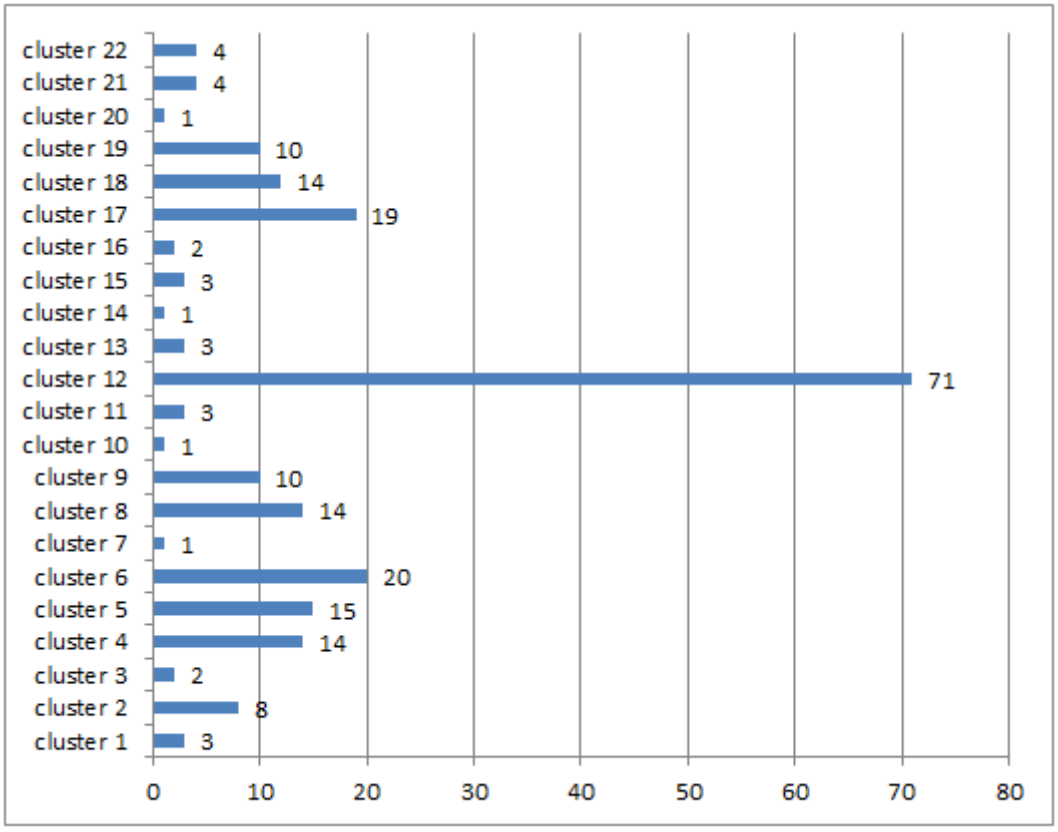
Percebe-se também, que nos clusters 1 e 8 e 3 e 4 os termos que mais ocorrem são os mesmos “foucault” e “cultura”, respectivamente. Isto pode ser explicado devido a escolha aleatório dos centroides, fazendo com que as publicações pertencentes à mesma série na distribuição original, seja distribuídas em clusters distintos.

A próxima seção apresenta os resultados para o último experimento.

4.5.3 Terceiro experimento

A Figura 35 apresenta o resultado do quarto (e último) experimento com a quantidade de publicações agrupadas para o número de 22 clusters.

Figura 34. Distribuição Gerada com 22 Clusters



Analisando-se esse último resultado, novamente é possível constatar um *cluster* que se sobressai em quantidade de publicações similares e *clusters* com quantidade de publicações similares muito próximas. No entanto, vale ressaltar a quantidade significativa de *clusters* com apenas uma publicação (*clusters* 7, 10, 14 e 20), ou seja, com apenas o centroide. Esse resultado mostra a importância da escolha do número de *clusters*, uma que pode refletir em agrupamentos com apenas um elemento.

A Tabela 9 apresenta os clusters identificados neste experimento e suas respectivas publicações.

Tabela 8. Estrutura de Cluster do Terceiro Experimento

Cluster	Publicações
Cluster 1	S8P64, S8P65, S8P66
Cluster 2	S2P9, S2P10, S2P11, S2P12, S2P13, S2P14, S2P15, S2P16
Cluster 3	S13P111, S13P112
Cluster 4	S17P156, S17P158, S17P159, S17P160, S17P161, S17P162, S17P163, S17P164, S17P165, S17P166, S17P167, S17P168,

	S17P171, S17P172
Cluster 5	S15P126, S19P185, S19P186, S19P187, S19P188, S19P189, S19P190, S19P191, S19P192, S19P193, S19P194, S19P195, S19P196, S19P197, S19P198
Cluster 6	S7P58, S10P79, S20P202, S20P205, S21P208, S21P208, S21P210, S21P211, S21P212, S21P213, S21P214, S21P215, S22P216, S22P217, S22P218, S22P219, S22P220, S22P221, S22P222, S22P223
Cluster 7	S7P59
Cluster 8	S16P139, S16P140, S16P141, S16P142, S16P143, S16P145, S16P146, S16P148, S16P150, S16P151, S16P152, S16P153, S16P154, S16P155
Cluster 9	S3P18, S3P19, S3P20, S3P21, S3P22, S3P23, S3P24, S12P100, S12P102, S12P104
Cluster 10	S17P169
Cluster 11	S20P200, S20P201, S20P204
Cluster 12	S1P1, S1P3, S1P4, S1P5, S1P6, S2P17, S4P25, S4P26, S4P27, S5P31, S6P50, S6P51, S6P52, S6P53, S6P54, S6P55, S6P56, S7P57, S7P60, S7P61, S7P62, S7P63, S9P71, S9P74, S9P76, S10P78, S10P80, S10P81, S10P82, S10P83, S12P101, S12P103, S12P105, S12P106, S13P107, S13P108, S13P109, S13P110, S13P113, S14P114, S14P115, S14P116, S14P117, S14P118, S14P119, S14P120, S14P121, S14P122, S14P123, S14P124, S15P125, S15P127, S15P128, S15P129, S15P130, S15P131, S15P132, S15P133, S15P134, S15P135, S16P136, S16P137, S16P138, S16P144, S16P147, S17P157, S17P170, S20P199, S20P203, S20P206, S20P207
Cluster 13	S1P2, S1P7, S1P8
Cluster 14	S16P149
Cluster 15	S4P28, S4P29, S4P30
Cluster 16	S18P174, S18P181
Cluster 17	S5P32, S5P33, S5P34, S5P35, S5P36, S5P37, S5P38, S5P39, S5P41, S5P42, S5P43, S5P44, S5P45, S5P46, S5P47, S5P48,

	S5P49, S11P93, S11P98
Cluster 18	S11P84, S11P85, S11P86, S11P87, S11P88, S11P89, S11P90, S11P91, S11P92, S11P94, S11P95, S11P96, S11P97, S11P99
Cluster 19	S18P173, S18P175, S18P176, S18P177, S18P178, S18P179, S18P180, S18P182, S18P183, S18P184
Cluster 20	S5P40
Cluster 21	S8P67, S8P68, S8P69, S8P70
Cluster 22	S9P72, S9P73, S9P75, S9P77

Neste experimento ocorreu de quatro *clusters* possuírem somente uma publicação, neste caso, somente o centroide. Isso pode ser explicado pelo fato do algoritmo ter selecionado aleatoriamente duas publicações com características muito similares para centroide. Dessa forma, as demais publicações pertencentes originalmente à mesma série e que possui forte relação de similaridade, estavam mais próximas de um dos centroides, fazendo com que o outro não consiga encontrar publicações com características similares com o qual esteja mais próximo.

Tabela 9. Termos mais Frequentes do Terceiro Experimento

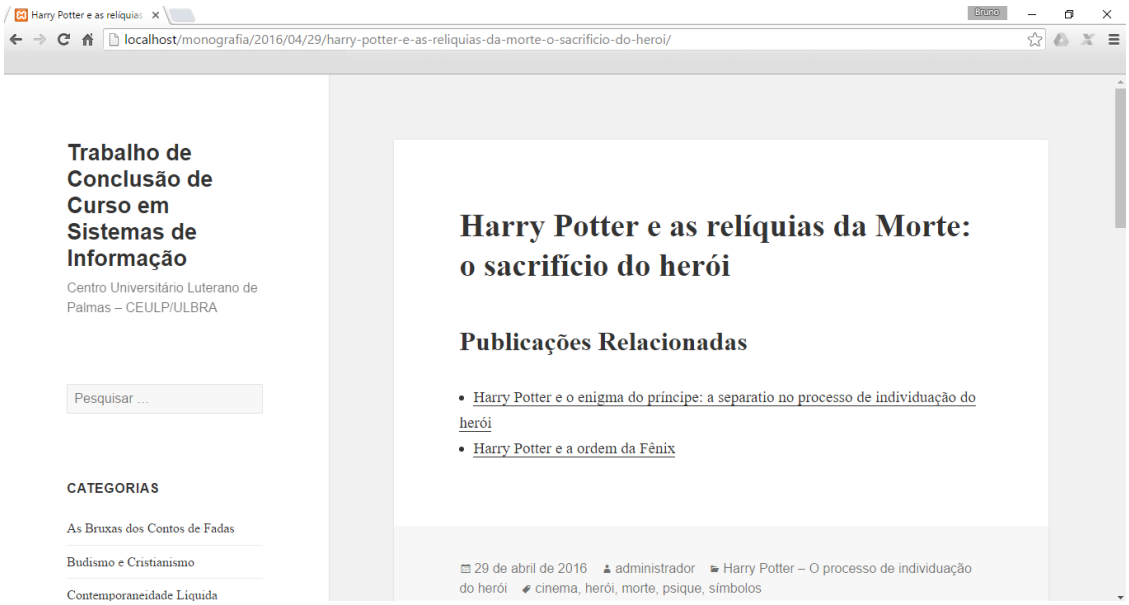
Cluster	**Termos mais frequentes (frequência de documentos)**
Cluster 1	cinema(3), psique(3), harry(2), porter(2)
Cluster 2	budismo(8), cristianismo(8), religião(8), filosofia(5), discurso(3), Deus(3), homem(2)
Cluster 3	psicodinâmica(2)
Cluster 4	cinema(12), Oscar2016(12), amor(2), criança(2), infância(2)
Cluster 5	Disney(15), princesas(15), estereótipos(12), feminino(8), fantasia(8), amor(3)
Cluster 6	religião(17), cultura(13), pecado(10), culpa(9), virtudes(8), vaidade(3), sexo(2), inveja(2)
Cluster 7	ciência(1), conhecimento(1), desafio(1), duvida(1), filosofia(1), iluminismo(1), liberdade(1), rebelião(1), teoria(1), oltaire(1)
Cluster 8	cinema(14), Oscar2015(14), herói(4), arquétipo(3), luto(2), viagem(2)
Cluster 9	modernidade(10), bauman(7), liquida(7), cartas(2), crônica(2),

	mulher(2)
Cluster 10	alterofobia(1), estrada(1), feminismo(1), furia(1), mad(1), max(1), Oscar2016(1)
Cluster 11	cultura(3), javaé(3), karajá(3), sofrimento(3), ação(2), fotografias(2), xambioá(2)
Cluster 12	cinema(25), Oscar2013(11), Oscar2014(9), extensão(5), filosofia(5), arquétipo(5), moderna(4), mulher(4), CEULP(3), universidade(3), suicídio(3), trabalho(3), tecnologia(3), luta(3), futebol(3), idolo(3), bruxa(3), conto de fadas(3), demônio(3), depressão(3), iny(2), karajá(2), jovens(2), Oscar(2), EUA(2), poder(2), CIA(2), escravidão(2), amor(2), sofrimento(2), mental(2), mito(2), saúde(2), comunidade(2), relacionamento(2), liberdade(2), filósofo(2), evolução(2), vida(2), terapia(2), brasil(2), copa(2), esporte(2), mãe(2)
Cluster 13	bruxas(3)
Cluster 14	alzheimer(1), cinema(1), cognição(1), doença(1), Oscar2015(1)
Cluster 15	anatomia(3), depressão(3)
Cluster 16	foucault(2), história(2), livro(2)
Cluster 17	mitologia(19), psicologia(18), analitica(16), deuses(12), grecia(7), jung(5), Deus(5), grega(5), arquétipo(3), grego(2), gregos(2), deusa(2)
Cluster 18	orixá(11), arquétipo(9), divindade(5), mitologia(5), africana(4), ambíguo(3), ciclos(3), movimentos(3), psicologia(3)
Cluster 19	foucault(10), livro(9), poder(4), sociedade(3), sujeito(2), saber(2)
Cluster 20	analitica(1), ares(1), deuses(1), grecia(1), mitologia(1), sicologia(1)
Cluster 21	harry potter(4), herói(2)
Cluster 22	tecnologia(4), futuro(3), humano(3)

Conforme é possível observar na Tabela 7, os *clusters* que possuem apenas um elemento agrupado (*clusters* 7, 10, 14 e 20), poderiam ter sido agrupados em outros clusters (*clusters* 6, 4, 8, 17, respectivamente) baseado na distribuição original das séries pelo *site* (EN)Cena. Este resultado novamente mostra a importância da escolha do número de *cluster*s e a de um critério de seleção para os centroides.

Na Figura 35 é apresentado o resultado do processo de recomendação para o último experimento.

Figura 35. Recomendação no Terceiro Experimento



A Figura 35 apresenta o resultado do processo de recomendação após visita a publicação “Harry Potter e as Relíquias da Morte: o sacrifício do herói”. Analisando a Tabela 8 é possível verificar a distribuição do cluster ao qual a publicação pertence e as publicações recomendadas. Como o cluster ao qual a publicação pertence possui apenas 3 publicações, o *plug-in* apresentou apenas 2 publicações como etapa de recomendação.

A seção seguinte apresenta mais algumas considerações sobre os experimentos.

4.5.4 Análise dos Resultados

Esta seção apresenta algumas considerações gerais sobre os quatro experimentos realizados, conclusões e sugestões são apresentados.

Analisando os quatros experimentos, é possível observar que todos os *clusters* agruparam em pelo menos uma publicação. Isto é possível porque a associação inicial (centróides) das publicações ao *cluster* é feita randomicamente, dessa forma pelo menos uma publicação é atribuída a um *cluster*. No entanto, em virtude da randomização dos centroides, analisando variações dos resultados, foi possível constatar que cada vez que o algoritmo é executado uma solução é encontrada, pois

para encontrar uma solução ótima de *clustering*, o algoritmo K-means é repetido várias vezes, começando de um *clustering* randômico inicial diferente.

Com os testes que foram realizados também se percebeu que nem sempre é possível gerar o mesmo conjunto de recomendações, pois existem alguns problemas inerentes ao algoritmo K-means, como a escolha do número de K incialmente.

Um ponto negativo constatado diz respeito à quantidade de *clusters* agrupados com elementos insuficientes que satisfaçam ao critério de recomendação, uma vez que apresentam poucos elementos agrupados similarmente. Esse resultado foi atingido mesmo em um número considerado pequeno de agrupamentos, conforme é possível observar na Figura 33, onde foi definido 8 agrupamento e como resultado foram agrupados *clusters* com 3 e 4 elementos.

Um resultado bastante peculiar, é que em todas as distribuições existem um *cluster* que apresenta um acentuado número de publicações. Vale salientar que este resultado conseguido está intimamente ligado à qualidade descrita nos marcadores. Se as publicações não possuem atributos em comum, é muito difícil para o algoritmo detectar semelhanças entre as publicações. Por fim, o algoritmo implementado atendeu plenamente aos objetivos propostos.

5 CONSIDERAÇÕES FINAIS

Neste trabalho foram abordados os conceitos sobre Recuperação da Informação, Sistemas de Recomendação, Técnicas de Recomendação. Além disso, foi estudada as APIs do *WordPress*, As compreensões destes conceitos foram de grande importância para que fosse possível a definição e o desenvolvimento do mecanismo de recomendação proposta nesse trabalho. Em relação às técnicas de recomendação, neste trabalho foi utilizada a técnica de recomendação baseado em conteúdo, implementando TF-IDF para o cálculo do peso na matriz de *clustering* e o algoritmo de clustering K-means para agrupamento das publicações similares.

O clustering K-means cria um conjunto de k clusters e distribui o conjunto de publicações entre esses *clusters* usando a similaridade entre os vetores de publicações (conjunto de marcadores) e os centroides dos *clusters*. A similaridade entre uma publicação e um centroide é calculada como o somatório de todos os vetores que contém publicações no cluster dividido pelo número de vetores de publicações*.* O funcionamento do K-means se dá pela melhor definição dos k centroides que melhor representem os dados do sistema. Neste projeto foi utilizada distância euclidiana entre os itens, calculada por meio dos marcadores que descrevem cada publicação. Ela assume a hipótese que os conteúdos acessados por grupos de usuários com perfis semelhantes possuem grande probabilidade de ser de interesse de um usuário individual pertencente a este grupo.

Após realizar a implementação do mecanismo de recomendação, optou-se por criar e utilizar uma pequena base de dados em ambiente local. Dessa forma, ao realizar os testes pode-se comprovar a funcionalidade do mecanismo. O mecanismo de recomendação implementado atendeu plenamente aos objetivos propostos, uma vez que os testes apresentados mostraram a aplicabilidade da solução proposta para o contexto da Plataforma *WordPress* para agrupamento de publicações similares e a sua recomendação.

Embora os resultados sejam encorajadores, eles não podem ser considerados conclusivos, uma vez que foram feitas com uma base de dados salvas em um servidor local (*localhost*). Assim, para trabalhos futuro, é considerada a implantação do *plug-in* em um ambiente real, incluindo dados *on-line.*

As publicações resultantes da execução do experimento foram processadas sem nenhuma espécie de pré-processamento. É interessante o enriquecimento do mecanismo de recomendação com várias técnicas de extração e processamento de dados, como *stemming* e remoção de *stopwords*. A taxa de acerto dos diferentes métodos de *clustering* executados para as situações sem e com o uso de stemming poderiam ser avaliadas e comparadas. Por considerar somente as *tags* existentes nas publicações, optou-se por não utilizar a etapa de pré-processamento.

Além disso, vale salientar que o algoritmo de agrupamento implementado é altamente dependente da escolha inicial do número de *cluster*. É interessante analisar se o valor de k influencia no desempenho no tempo necessário para o cálculo da filtragem baseado em conteúdo, ou seja, implica em um menor tempo de resposta da geração das recomendações.

Outro ponto importante como trabalhos futuros é a criação de outras técnicas de recomendação como técnicas baseadas em Filtragem Colaborativa, avaliando a qualidade das recomendações geradas no contexto da Plataforma *WordPress*.

6 REFERÊNCIAS

BAEZA-YATES, R; RIBEIRO-NETO, B. **Modern Information Retrieval**. 1st Edition. Harlow: Addison Wesley, 1999. Disponível em: < ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf> Acesso em: 14 de jul. 2015.

BEZERRA, B. L. D. **Estudo de Algoritmos de Filtragem de Informação Baseados em Conteúdo**. 2006. 45p. TCC (Graduação em Ciência da Computação) - Universidade Federal de Pernambuco, Pernambuco.

BORGES, D. M. **Estudo de Técnicas de Recomendações Automáticas de Produtos.** 2010. 85p. Relatório de Estágio (Graduação em Sistemas de Informação) – Centro Universitário Luterano de Palmas, Palmas. Disponível em: <http://followscience.com/content/680/estudo-de-tecnicas-de-recomendacoes-automaticas-de-produtos> Acesso em: 26 de jul. 2015.

CARDOSO, O. N. P. Recuperação de Informação. INFOCOMP: Revista de Computação da UFLA, Lavras, v. 1, 2000. Disponível em: <www.dcc.ufla.br/infocomp/artigos/v2. 1/art07.pdf>. Acesso em: 29 de jul. 2015.

CASTRO, S. S. G. F.; BARBOSA, T. M. **RAPIDos**: Recomendação Automática de Produtos Interessantes em Dispositivos Móveis. 2009. 61p. TCC (Graduação em Ciência da Computação) – Universidade de Brasília, Brasília. Disponível em: em: <http://monografias.cic.unb.br/dspace/bitstream/123456789/191/1/Monografia_Saulo_Tiago_FINAL.pdf>. Acesso em: 17 de set. 2015.

CARVALHO, V. Como Criar *Plugins* para *WordPress*. Escola Web. 2012. Disponível em: <http://www.escolawp.com/2012/01/como-criar-plugins-para-wordpress-parte-i/> Acesso em: 3 de fev. 2016.

CAZELLA, S.C. NUNES, M.A.S.N. REATEGUI, E.B. **A ciência da opinião:** estado da arte em sistemas de recomendação. In: Anais do XXX Congresso da SBC Jornada de Atualização da Informática, 2010. Disponível em: <http://www.do.ufgd.edu.br/WillianAmorim/TAIC022013_arquivos/Artigo2.pdf> Acesso em: 28 de jul. 2015.

COSTA,E. AGUIAR, J. MAGALHÃES,J. **Sistemas de Recomendação de Recursos Educacionais:** conceitos, técnicas e aplicações. In: II Jornada de Atualização em Informática na Educação (JAIE). II Congresso Brasileiro de Informática na Educação (CBIE), páginas 57-78, 2013.

CORREIA, L. A; **Um Sistema de Recomendação de Promoções Baseado em Posts no Twitter**. 2011. 42p. TCC (Graduação em Ciência da Computação) – Universidade Federal de Pernambuco, Recife.

DEPPLER, F.D. **Um Modelo para Recuperação e Busca de Informação Baseado em Ontologia e no Círculo Hermenêutico.** 2008. 135p. Tese (Doutorado em Engenharia e Gestão do Conhecimento) - Universidade Federal de Santa Catarina, Florianópolis.

_____, F.D. TODESCO, J. L. GONÇALVES, A. SELL, D. MORALES, A.B.T. PACHECO, R.C.S. **Uma arquitetura para recuperação de informação aplicada ao processo de cooperação universidade-empresa**. 2005. Disponível em: <http://www.fbeppler.com/papers/UmaAquiteturaRecuperacaoInformacaoAplicadaProcessoCooperacaoUniversidadeEmpresa.pdf> Acesso em: 16 de jul. 2015

(En)Cena. Disponível em: <http://encenasaudemental.net/> Acesso em: 12 de fev. 2016.

FERRERO, C. A. **Algoritmo KNN para Previsão de Dados Temporais:** funções de previsão e critérios de seleção de vizinhos próximo aplicados a variáveis ambientais em limnologia. 2009. 129p. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) Instituto de Ciências Matemáticas e de Computação, São Carlos.

FONSECA, F.C.S. BELTRAME, W.A.R. **Aplicações Práticas dos Algoritmos de Clusterização K-Means e Bisecting K-Means**. UFES, 2010. Disponível em: <http://www.inf.ufes.br/~claudine/courses/paa10/seminarios/seminario4.pdf> Acesso em: 28 de jul. 2015.

GREENGRASS, E. Information Retrieval: A Survey. 2000. (TR-R52-008-001)

GIORDANO, R.B. **Da Necessidade ao conhecimento: Recuperação da Informação na Web em Ciência da Informação.** 2011. 145p. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Jain, A. K., Murty, M. N., and Flynn, P. J. **Data clustering:** a review. ACM Computing Surveys, Vol. 31, No. 3, September 1999.

MEDEIROS, I. R. G. **Estudo sobre sistemas de recomendação colaborativos**. 2013. 30p. TCC (Graduação em Ciência da Computação) - Universidade Federal de Pernambuco, Pernambuco. Disponível em: <http://www.cin.ufpe.br/~tg/2012-2/irgm.pdf> Acesso em: 20 de jul. 2015.

OLIVEIRA, F.L. **Clusterização de Consultas em um Modelo de Ordenação Web Baseado na Relevância por Tempo em Domínio Aberto.** 2005. 78p. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis

REHMAN M. **Comparison of Density-based Clustering Algorithms**, Lahore College for Women University Lahore. 2006. Disponível em: < https://www.researchgate.net/publication/242219043_COMPARISON_OF_DENSITY-BASED_CLUSTERING_ALGORITHMS > Acesso em: 30 de jul. 2015.

ROUSSEAU, B., BROWNE, P., MALONE, P. FOSTER, P., MENDIS, V. **Personalised resource discovery searching over multiple repository types:** Using user and information provider profiling. In ICEIS (5), 2004. Pag. 35–43.

SALTON, G.; McGILL, M. J. **Introduction to modern information retrieval**. New York, McGraw-Hill, 1983. Disponível em: <http://lyle.smu.edu/~mhd/8337sp07/salton.pdf>. Acessado em: 29 de set. 2015

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Perspectiva em Ciência da Informação**, Belo Horizonte, v.1, n.1, p.41-62, jan./jun. 1996.

_____T. Information Science. Journal of The American Society for Information Science, v.50, n.12. p.1051-1063. 1999.

SILLA JR., C. N.; KAESTNER, C. A. A. **Estudo de Métodos Automáticos para Sumarização de Textos**. In: Simpósio de Tecnologias de Documentos, 2002, São Paulo. Anais do STD 2002. São Paulo: ITS, 2002, v. 1, p. 45-49. Disponível em: <http://sites.google.com/site/carlossillajr/files/2002-STD.pdf?attredirects=0>. Acesso em: 19 de fev. 2016.

STAKOVIAK, F.H.M. **Implantação de um Mecanismo de Enriquecimento do Perfil do Usuário e Recomendação de Trabalhos científicos para o Konnen.** 2011. 74p. TCC (Graduação em Ciência da Computação) - Centro Universitário Luterano de Palmas, Palmas.

STEINBACH, M.; KARYPIS, G.; KUMAR, V. A **Comparison of Document Clustering Techniques.** In KDD Workshop on TextMining, 2000. Disponível em: <http://rakaposhi.eas.asu.edu/cse494/notes/clustering-doccluster.pdf>. Acessado em: 29 de jul. 2015.

TAVARES, B.M. **Sistema de Recomendação de e-Learning**. 2012, 76p. Dissertação (Mestrado em Engenharia Informática, Área de Especialização em Arquiteturas, Sistemas e Redes) - Instituto Superior de Engenharia do Porto, Porto.

Tele Sintese - Portal de Telecomunicações, Internet e TICs. Até dezembro, mundo terá 3 bilhões de internautas. 2014. Disponível em: <http://www.telesintese.com.br/35910/> Acesso em: 12 jun. 2015.

TSUJI, G.K. Kamaura, L. T. **Integrando Recuperação de Informação em Banco de Dados com *Hibernate Search***. 36 P. TCC (Graduação em Ciência da Computação) - Instituto de Matemática e Estatística, São Paulo.

VENSON, E. **Um Modelo de Sistema de Recomendação Baseado em Filtragem Colaborativa e Correlação de Itens para Personalização no Comércio Eletrônico.** 2002. 132p. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis

VIEIRA, J. A. Excesso de Informação e Correria: ameaça à saúde física e mental. 2012. Disponível em: <http://simoesvieira.adv.br/artigo/excesso-de-informacoes-e-correria-ameaca-a-saude-fisica-e-mental.html> Acesso em: 17 jun. 2015.

WIVES, L.K. **Utilizando Conceitos como Descritores de Textos para o Processo de Identificação de Conglomerados (*clustering*) de Documentos**. 2004. 136p. Tese (Doutorado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul, Porto Alegre.

WORDPRESS. Disponível em <http://br.wordpress.org/> Acesso em: 9 de fev. 2016.

ANEXOS

Publicações (título e marcadores) utilizadas nos experimentos extraídas do Portal (EN)Cena.

Seção: As Bruxas dos Contos de Fadas

S1P1: Rainha de copas: da albedo à rubedo, uma análise alquímica da função sentimento.
albedo,alice,alquímica,carroll,copas,lewis,rainha,rubedo.

S1P2: Os cisnes selvagens e o animus feminino.
animus,bruxas,cisnes,feminino.

S1P3: A Bruxa Morgana e o matriarcado pagão
arquétipo,bruxa,matriarcado,morgana,paganismo.

S1P4: João e Maria: aspectos simbólicos do inconsciente
arquétipo,conto de fadas,joão,mãe,maria

S1P5: Mãe Gothel e a busca pela beleza eterna
beleza eterna,bruxa,conto de fadas,filme,Rapunzel

S1P6: Úrsula e o arquétipo da Mãe Terrível
arquétipo,bruxa,conto de fadas,mãe,poder e vingança

S1P7: A Rainha Má de Branca de Neve e a inveja
bruxas,conto de fadas,inveja,mãe

S1P8: As bruxas dos contos de fadas: aspectos sombrios da alma feminina
alma,bruxas,contos,fadas,feminina

Série: Budismo e Cristianismo

S2P9: É possível um diálogo inter-religioso entre Budismo e Cristianismo?
aproximação,budismo,cristianismo,discurso,discurso,discurso,religião

S2P10: O que Budismo e Cristianismo têm de diferente, pela análise de Lubac e Usarski
amor, budismo, compaixao, conhecimento, cristianismo, Deus, filosofia, homem, libertação, religião.

S2P11: Aproximações e distanciamentos: Além da insubstancialidade
budismo,cristianismo,Deus,existência,filosofia,religião,verdade

S2P12: A genealogia do diálogo inter-religioso entre Budismo e Cristianismo
budismo,cristianismo,cultura,genealogia,religião,sunyata,teologia,transcendencia

S2P13: Eckhart é um elo com o Oriente
budismo,cristianismo,Deus,filosofia,homem,humanismo,iluminismo,religião,sagrado

S2P14: Pontos de contato filosófico entre o Cristianismo e o Budismo
budismo,cristianismo,filosofia,imanencia,religião

S2P15: Budismo e Cristianismo: Há mais em comum do que se possa imaginar
budismo,cristianismo,religião,tradições

S2P16: Imanência e Transcendência: chaves de leitura entre o Budismo e o Cristianismo
budismo,cristianismo,dialogo,filosofia,religião

S2P17: Notas
conceitos,notas

Série: Contemporaneidade Líquida

S3P18: Outras cartas do mundo líquido moderno
bauman,cartas,liquida,modernidade

S3P19: Algumas cartas para o mundo líquido moderno
bauman,cartas,liquida,modernidade

S3P20: Vida Líquida: consumo, velocidade e lixo na era da incerteza
bauman,liquida,modernidade,vida

S3P21: O medo líquido em Bauman: leituras possíveis
bauman,liquida,medo,modernidade

S3P22: Amor Líquido: a problemática das relações amorosas e dos vínculos familiares na literatura contemporânea
amor,bauman,liquida,modernidade

S3P23: Modernidade Líquida: andando sobre uma fina camada de gelo
bauman,liquida,modernidade,sociedade

S3P24: Diálogos contemporâneos impertinentes com/de Bauman
bauman,contemporaneidade,liquida,modernidade

Série: Demônio em Fuga

S4P25: Demônio em Fuga: a luta continua
demônio,depressão,luta,tratamento

S4P26: Uma anatomia da cura
demônio,depressão,luta,vida

S4P27: Uma anatomia da depressão
demônio,depressão,terapia,vida

S4P28: Uma anatomia da dor
alma,anatomia,byron,cura,depressão

S4P29: Uma anatomia do ódio
agorafobia,anatomia,depressão,morte,toc,vida

S4P30: Uma anatomia da destruição
anatomia,bipolar,depressão,destruição,nascer,psicopatologia

Série: Deuses Gregos

S5P31: Pã e o arquétipo dos instintos primitivos
arquétipo,instintos,primitivo

S5P32: Hades e o arquétipo da força que impulsiona o crescimento
analitica,arquétipo,Deus,hades,mitologia,psicologia

S5P33: Demeter e a representação do instinto maternal
analitica,demeter,Deus,grego,mitologia,psicologia

S5P34: Hécate – a deusa ctônica
analitica,deuses,grecia,hecate,mitologia,psicologia

S5P35: Dionísio – o deus do vinho, da loucura, do êxtase e da tragédia
analitica,Deus,dionisio,grega,mitologia,psicologia

S5P36: Perséfone – a deusa virgem e rainha do submundo
analitica,Deus,grego,mitologia,persefone,psicologia

S5P37: Zeus – O Senhor do Olimpo
analitica,deuses,grecia,mitologia,psicologia,zeus

S5P38: Apolo e a sombra da distância emocional
analitica,apolo,euses,grecia,mitologia,psicologia

S5P39: Hera – A grande mãe

analitica,deuses,gregos,hera,mitologia,psicologia

S5P40: Ares e o arquétipo da força física
analitica,ares,deuses,grecia,mitologia,psicologia

S5P41: Poseidon – o arquétipo incoercível das fortes emoções
analitica,deuses,grecia,mitologia,poseidon,psicologia

S5P42: Artemis – a deusa dos instintos
analitica,artemis,deusa,deuses,grecia,mitologia,psicologia

S5P43: Eros e o desejo incoercível dos sentidos
analitica,cupido,deuses,eros,grecia,mitologia,psicologia

S5P44: Héstia e o fogo sagrado da purificação
analitica,deuses,grega,hestia,junguiana,mitologia,psicologia

S5P45: Hermes: o guia das almas
analitica,apsicologia,asas,deuses,grega,hermes,mercúrio,mitologia,vento

S5P46: Atena e o arquétipo da sabedoria
analitica,arquétipo,atena,deuses,gregos,jung,minerva,mitologia,psicologia

S5P47: Cronos e a passagem do tempo
analitica,cronos,Deus,deuses,goya,grega,jung,mitologia,psicologia,tempo

S5P48: Afrodite e o arquétipo do amor
afredite,amor,arquétipo,deusa,grecia,jung,mitologia,psicologia,venus

S5P49: As lições de Prometeu
deuses,grega,herói,mitologia,prometeu,psicologia,semideus

Série: Encena em Campo

S6P50: #vaitercopa
brasil,copa,envorganhado,esperança,promessas

S6P51: Don Diego Armando Maradona, um gigante entre os gigantes
argentino,esporte,futebol,idolo,talento

S6P52: Redemocratização “lenta, gradual e segura” e neoliberalismo: aquela como farsa essa como tragédia
democracia,ditadura,farsa,neoliberalismo,transição

S6P53: Garrincha: o “anjo de pernas tortas”

agilidade,esporte,futebol,idolo,pernas

S6P54: A seleção brasileira, Médici e os anos de chumbo
brasil,golpe,imagem,propaganda,seleção brasileira

S6P55: Futebol e Copa – quando a paixão vira negócio o povo não entra no jogo
capitalismo,copa,economia,negocio,sociedade

S6P56: Pelé: o jogador
brasileiro,futebol,idolo,rei do futebol,velocidade

Série: Fragmentos do Saber

S7P57: René Descartes: penso, logo existo
existência,filosofia,Rene Descartes

S7P58: Kierkegaard – A severidade do Deus que promete e cumpre
Deus,divino,fé,filosofia,filósofo,Kierkegaard,religião,severidade

S7P59: Voltaire – Toda teoria pode ser desafiada
ceticismo,ciência,conhecimento,desafio,duvida,filosofia,iluminismo,liberdade,rebelião,teoria,Voltaire

S7P60: Baruch Espinosa e o início da era secular
baruch,espinosa,filosofia,liberdade

S7P61: John Stuart Mill: um estudo das motivações e das justificativas das ações
filosofia,filósofo,John Stuart Mill

S7P62: Um recado na geladeira deixado por Nietzsche
filosofia,filósofo,Nietzsche

S7P63: "Nomes da Filosofia" – Os pensadores por trás das grandes ideias
filosofia,filósofos,história,pensadores

Série: Harry Potter – O processo de individuação do herói

S8P64: Harry Potter e as relíquias da Morte: o sacrifício do herói
cinema,herói,morte,psique,símbolos

S8P65: Harry Potter e o enigma do príncipe: a separatio no processo de individuação do herói
cinema,harry,inconsciente,porter,psique

S8P66: Harry Potter e a ordem da Fênix
cinema,fênix,harry,porter,psique

S8P67: Harry Potter e o Cálice de fogo: a opus alquímica
alquimia,cálice de fogo,harry potter,tribruxo

S8P68: Harry Potter e o prisioneiro de Azkaban: e a projeção do complexo paterno
complexo,harry potter,prisioneiro de Azkaban,projeção

S8P69: Harry Potter e a Câmara Secreta – a transformação do herói
bullying,câmara secreta,harry potter,herói

S8P70: Harry Potter e a Pedra Filosofal: a jornada do herói
harry potter,herói,pedra filosofal,voldemort

Série: Humano Demasiado Tecnológico

S9P71: Drones: entre evolução tecnológica, poder e bom senso
drone,evolução,tecnologia

S9P72: Mundo Wired: conexão, velocidade e virtualização
conexão,futuro,tecnologia,virtualização

S9P73: Eu, um robô? – A codificação da empatia
empatia,futuro,humano,robô,tecnologia

S9P74: O que você escreve no Twitter define quem você é?
Mídia,personalidade,ser importante,softwares,twitter

S9P75: Bem-vindo ao Futuro, Humano!
futuro,humano,tecnologia

S9P76: Dados, dados e mais dados: o fenômeno Big Data
dados,evolução,informação,tecnologia

S9P77: Tecnologia, Humanidade e (R)Evoluções
humano,inovação,tecnologia

Série: La Fora

S10P78: (En)Cena – A Saúde Mental em Movimento
(en)cena,CEULP,comunidade,extensão

S10P79: Cofo da leitura e escrita na escola indígena
comunidade,cultura,extensão,indígena,leitura

S10P80: EXPRO – Exposição das Profissões
CEULP,EXPRO,extensão,profissão,universidade

S10P81: Akádemo – uma lição de vida
akademo,extensão,trote,universidade

S10P82: Fisioterapia Aquática na APAE de Palmas – TO
APAE,comunidade,extensão,fisioterapia,Palmas

S10P83: TERRAQUARIUM – Educação e Meio Ambiente
ambiental,CEULP,extensão

Série: Mitologia Africana

S11P84: Oxum e o arquétipo da feminilidade
amor,arquétipo,divindade

S11P85: Omulú e o arquétipo do curador ferido
arquétipo,divindade,ritual

S11P86: Oxalá e o símbolo da criação
criação,criador,orixá

S11P87: Oxumaré: o símbolo da continuidade e permanência
ambíguo,arquétipo,ciclos,movimentos,orixá

S11P88: Oxumaré: o símbolo da continuidade e permanência
ambíguo,arquétipo,ciclos,movimentos,orixá

S11P89: Oxumaré: o símbolo da continuidade e permanência
ambíguo,arquétipo,ciclos,movimentos,orixá

S11P90: Obá – deusa do amor e da paixão incontrolável
divindade,mulher,orixá

S11P91: Ewá e o arquétipo da castidade
castidade,deusa,divindade,feminina,orixá

S11P92: Nanã – a mãe terra
africana,arquétipo,mãe,mitologia,nana,orixá,psicologia

S11P93: Xangô – o fogo que rasga o céu
africana,analitica,jung,mitologia,psicologia,xango

S11P94: Ossain e a transformação da realidade
divindade,natureza,orixá

S11P95: Iansã – Senhora dos relâmpagos e das tempestades
africana,iansa,jung,mitologia,orixá,psicologia

S11P96: Exú – o guardião dos caminhos
arquétipo,exu,mitologia,orixá,psicologia

S11P97: Oxóssi e o arquétipo da liberdade
africana,analitica,arquétipo,candoble,mitologia,orixá,oxossi

S11P98: Iemanjá: rainha das águas
africanos,deuses,iemanja,jung,mãe,mitologia,odoia,orixá,psicologia,religião

S11P99: Ogum e o arquétipo da coragem
africana,arquétipo,mitologia,ogum

Série: Mulheres Modernas

S12P100: Diário de uma Mulher Moderna
feliz,independente,modernidade,mulheres

S12P101: Manhã de setembro de uma Mulher Moderna
moderna,mulher

S12P102: Mulheres Modernas – O preço da modernidade
crônica,independência,liberdade,moderna,modernidade,mulher,separação,sonho

S12P103: Mulheres Modernas: Relacionamento Virtual, por que não?
afetividade,moderna,mulher,relacionamento,virtual

S12P104: Checklist de Mulheres Modernas – Porque modernidade exige critérios
amizade,crônica,modernidade,mulher,relacionamentos

S12P105: Páginas da vida de uma Mulher Moderna – O Jantar
amigas,amizade,crônica,filhos,moderna,mulheres,solidão

S12P106: Vida de Mulher Moderna – A descoberta
independente,liberdade,moderna,mulher

Série: Mundo do Trabalho e Sofrimento Psiquco

S13P107: Ser bancário: o sujeito, o sofrimento e seus destinos
bancário,destino,sofrimento,sujeito,trabalho

S13P108: “Tu trabalha também ou só dá aula?”: Uma reflexão sobre o trabalho de docentes universitários
contextualização,metodologia,trabalho,universidade

S13P109: Loucos ou heróis? Educadores sociais e adolescentes em situação de rua
adolescente,educadores,heróis,loucos,ressignificar,rua

S13P110: Mito de Sísifo ou dos trabalhadores de saúde mental
mental,mito,saúde,sísifo,trabalhadores

S13P111: Clínica Psicodinâmica do Trabalho com a Unidade de Operações Aéreas do Detran
clínica,detran,psicodinâmica,psicopatologia,sofrimento,trabalho

S13P112: Clínica Psicodinâmica da Cooperação com catadores de materiais recicláveis
ascampa,catadores,psicodinâmica,recicláveltrabalho

S13P113: Mundo do trabalho e sofrimento psíquico
mundo,psíquico,sofrimento,trabalho

Série: Oscar 2013

S14P114: A Caça
alienação,cinema,estigma,Oscar2013

S14P115: “As sessões” e o sexo terapêutico
cinema,Oscar2013,sexo,terapia

S14P116: Lincoln
cinema,EUA,Lincoln,Oscar2013,poder

S14P117: Argo
CIA,cinema,Oscar2013,poder

S14P118: Amour
amour,cinema,morte,Oscar2013

S14P119: Django Livre
cinema,escravidão,Oscar2013

S14P120: Os Miseráveis
cinema,movimentos,Oscar2013,revolução

S14P121: Indomável Sonhadora
cinema,imaginação,Oscar2013

S14P122: A hora mais escura
CIA,cinema,guerra,Oscar2013

S14P123: O Lado Bom da Vida
bipolar,cinema,doente mental,Oscar2013

S14P124: As aventuras de Pi
cinema,naufrágio,Oscar2013,sobrevivência

Série: Oscar 2014

S15P125: HER: a incompletude palatável
amor,bauman,cinema,contemporaneo,cuidado,her,lacos,scarlettjohansson,sensibilidade,tecnologia

S15P126: Frozen – Uma aventura congelante
animação,cinema,Disney,Oscar2014,princesas

S15P127: Inside Llewyn Davis: a propósito de um Ulisses sem Odisséia
cinema,odisséia,Oscar2014

S15P128: Da pompa à decadência, Blue Jasmine mostra a dificuldade de se encarar as mudanças
cinema,decadência,Oscar2014,vulnerabilidade

S15P129: Nebraska: delusão monocromática numa América sem "maquiagem"
cinema,ilusão,Oscar2014,trajetória

S15P130: Philomena: a relação paradoxal da mulher na religião
cinema,mulher,Oscar2014,religião

S15P131: Clube de Compras Dallas: na tragédia, a mudança
AIDS,cinema,Oscar2014,tragédia

S15P132: Capitão Phillips
cinema,claustrofobia,Oscar2014,tensão

S15P133: O Lobo de Wall Street: poder, vício e manipulação
cinema,manipulação,Oscar2014,vicío

S15P134: Before Midnight – Antes da Meia-Noite

cinema,Oscar2014,relacionamento

S15P135: Gravidade: um universo de silêncio e solidão
cinema,Oscar2014,silêncio,universo

Série: Oscar 2015

S16P136: Selma: é preciso acreditar, agir e seguir em frente!
cinema,escravidão,luta,Oscar,personagem

S16P137: A vingança no micro-ondas: Relatos Selvagens
fragilidade,realidade,reflexão

S16P138: Dois dias, uma noite – "O inferno são os outros"
cinema,inferno,Oscar,outros

S16P139: Whiplash e a sofrida (e instigante) busca pela perfeição
busca,cinema,Oscar2015,perfeição,sonho

S16P140: Guardiões da Galáxia: a jornada do herói
arquétipos,cinema,equipe,herói,Oscar2015

S16P141: Como Treinar seu Dragão 2: o despertar da consciência
cinema,consciência,dragão,Oscar2015,treino

S16P142: Birdman ou A Inesperada Virtude da Ignorância
cinema,Oscar2015,relação,self,vida

S16P143: Operação Big Hero: uma perspectiva de elaboração do luto na adolescência
arquétipo,cinema,herói,luto,Oscar2015

S16P144: O conto da Princesa Kaguya
arquétipo,conto,mito

S16P145: Sniper Americano e o (ainda) insuperável "mito do herói"
arquétipo,cinema,herói,mito,Oscar2015

S16P146: Wild – Livre: Uma Jornada de Autoconhecimento
autoconhecimento,cinema,luto,Oscar2015,viagem

S16P147: O Hobbit – A batalha dos cinco exércitos
arquétipos,inconsciente,simbolismo

S16P148: Amor e entrelaçamento quântico no filme "Interestelar"

amor,ciência,cinema,Oscar2015,viagem

S16P149: Para sempre Alice – Alzheimer e a Arte de Perder
alzheimer,cinema,cognição,doença,Oscar2015

S16P150: Caminhos da Floresta e a psicologia dos contos de fadas
arquétipo,cinema,conto de fadas,musical,Oscar2015

S16P151: Alan Turing e O Jogo da Imitação – O que Significa Ser Humano?
cinema,guerra,herói,Oscar2015,tecnologia

S16P152: Boyhood – Da Infância à Juventude
cinema,crescimento,desenvolvimento,Oscar2015,tempo,transformações

S16P153: "O Grande Hotel Budapeste": ode à amizade e à resiliência"
amizade,cinema,Oscar2015,resiliência,transformação

S16P154: "Garota Exemplar" e a tragédia como delineadora da vida
bauman,cinema,liquidez,Oscar2015,persona,tragédia

S16P155: Malévola e a redenção do feminino ferido
cinema,feminino,Malévola,Oscar2015

Série: Oscar 2016

S17P156: "A Grande Aposta" mostra de forma cômica a crise econômica de 2008
aposta,cinema,crise,economica,EUA,Oscar2016

S17P157: "Ponte dos Espiões": o medo como mecanismo de controle
cinema,espioes,EUA,oscar,tomhanks

S17P158: Trumbo: a indústria cinematográfica desconstruída
ator,cinema,espionagem,guerra,Oscar2016,roteiro,trumbo

S17P159: "Carol" e o caminho da completude feminina
amor,Cate-Blanchett, cinema, feminino, gênero, homoafetividade, homofobia, mulher, Oscar2016, Rooney-Mara, sexualidade

S17P160: O menino e o mundo: a distopia em suas possibilidades
animação,brasil,cinema,crescimento,criança,criança,favela,infância,pobreza,transformação

S17P161: The Revenant – O Regresso à Natureza Selvagem
cinema,DiCaprio,Oscar2016,regresso

S17P162: “A Garota Dinamarquesa” e o fim da era das certezas
androginia,casamento,cinema,comportamento,Eddie-Redmayne,
gênero,Oscar2016,sexo,sexualidade,teoria queer,transexualidade

S17P163: Brooklin – O amor e o processo de individuação
amor,cinema,drama,família,Oscar2016,romance

S17P164: Spotlight – Segredos Revelados: quando a verdade se oculta na manipulação da fé
cinema,igreja,Oscar2016,spotlight

S17P165: Star Wars – O Despertar da Força: o herói solar
cinema,episodio8,hansolo,kylo,lucasfilme,Oscar2016,prioncesalea. ficcao,starwars

S17P166: Ex Machina: a senciência da criação
android,cinema,ficção,frankenstein,informatica,inteligência,tecnologia,Oscar2016

S17P167: O Quarto de Jack: quando a resiliência cria o impossível
cinema,infância,Oscar2016,resiliência

S17P168: Perdido em Marte: há limites para a expansão humana?
cinema,essência,marte,Oscar2016,superação

S17P169: “Mad Max: Estrada da Fúria” – Aridez apocalíptica é cenário para “feminismos” e alterofobia
alterofobia,estrada,feminismo,furia,mad,max,Oscar2016

S17P170: 50 tons de cinza – porque o óbvio passa despercebido
amor,cinema,comportamento,gênero,relação

S17P171: Cinderela e o processo de individuação nos contos de fadas
animus,conto de fadas,Oscar2016,processo de individuação

S17P172: Divertida Mente: Quem disse que é fácil crescer?
cinema,crescer,divertida,mente,Oscar2016

Série: Poder Subjetividade Saber

S18P173: Vigiar e Punir – história da violência nas prisões
controle,foucault,livro,prisoes,punir,sociedade,vigiar

S18P174: História da Loucura (1961) de Michel Foucault
foucault,história,livro,loucura,psiquiatria,Reforma

S18P175: Masculinidades nas malhas do biopoder – A emergência da Política de Atenção Integral à Saúde do Homem
bipoder,foucault,homem,integral,livro,política,sujeito

S18P176: O poder psiquiátrico de Michel Foucault – para interrogar o presente
anormal,foucault,livro,patologico,poder,psiquiatria

S18P177: As palavras e As Coisas – os Conhecimentos, as Pessoas e Suas Ordens
coisas,conhecimento,foucault,livro,ordens,palavras,pessoas

S18P178: Doença Mental e Personalidade – o ser diante da patologia ou a patologia inerente ao sujeito
foucault,livro,patologia,personalidade,saúde,sujeito

S18P179: Os Anormais
anormais,foucault,indivíduo,livro,periculosidade,sociedade

S18P180: O Nascimento da Clínica em Foucault – um Poder-Saber sobre a vida
clínica,foucault,livro,nascimento,poder,saber,vida

S18P181: História da sexualidade I – A vontade de saber
foucault,história,homossexualidade,livro,sexo,>sexualidade

S18P182: Em Defesa da Sociedade – Estudos de Biopolítica
biopolítica,defesa,estudos,foucault,livro,sociedade

S18P183: A ordem do discurso
discurso,foucault,livro,ordem,poder

S18P184: Poder – Subjetividade – Saber: diálogos com Foucault
foucault,poder,saber,subjetividade

Série: Princesas Disney

S19P185: Princesas Disney – O Amor como estatuto privilegiado no universo feminino
amor,Disney,estereótipos,feminino,princesas

S19P186: Anna e Elsa: tempestades de gelo, descobertas e diferenças
Disney,estereótipos,feminino,frozen,princesas

S19P187: Merida – Uma Princesa Diferente?
Disney,estereótipos,feminino,Merida,princesas

S19P188: Rapunzel – Em busca do seu lugar no mundo
Disney,fantasia,feminino,princesas

S19P189: Tiana: a magia e a realidade da quebra de preconceitos
amor,Disney,estereótipos,fantasia,princesas

S19P190: Mulan – a ruptura de estereótipos e a polissemia feminina
Disney,estereótipos,feminino,princesas

S19P191: Pocahontas: livre e independente em busca de seu caminho
Disney,estereótipos,fantasia,feminino,princesas

S19P192: Princesa Jasmine: entre seguir e transgredir as normas
Disney,estereótipos,fantasia,feminino,princesas

S19P193: Conhecimento e imaginação no universo feminino retratado em "A Bela e A Fera"
comportamento,conhecimento,Disney,feminino,princesas

S19P194: Entre conchas e algas: Ariel, a princesa ruiva pseudomoderna
ariel,Disney,estereótipos,fantasia,princesas

S19P195: Bela Adormecida: entrega, fragilidade e espera
aurora,Disney,estereótipos,fantasia,princesas

S19P196: Cinderela: estereótipo feminino no contexto do casamento/amor romântico
amor,cinderela,Disney,estereótipos,princesas

S19P197: Branca de Neve e as ilusões femininas no terceiro milênio
Disney,estereótipos,fantasia,princesas

S19P198: Princesas Disney: estereótipos e o universo feminino
consumo,Disney,estereótipos,fantasia,princesas

Série: Saúde Mental Indígena

S20P199: Educação indígena como estratégia de prevenção do suicídio
educação indígena,Hany,iny,karajá,suicídio

S20P200: Ação no Povo Karajá-Xambioá
ação,cultura,fotografias,javaé,karajá,sofrimento,xambioá

S20P201: Ação no Povo Karajá-Xambioá - parte 2
ação,cultura,fotografias,javaé,karajá,sofrimento,xambioá

S20P202: Ação no Povo Karajá-Xambioá - parte 3
cultura,direitos,Karajá-Xambioá,sofrimentos

S20P203: Do etnocentrismo ao pluriculturalismo: a saúde mental do indígena numa perspectiva biopsicossocial
aldeias,ilha do bananal,indígenas,iny,jovens,karajá,suicídio

S20P204: Como se faz saúde mental indígena?
cultura,indígena,iny,javaé,karajá,saúde mental,sofrimento,suicídio,território

S20P205: Ação no Povo Apinajé
Apinajé,cultura,fotografias,indígenas

S20P206: Ação no Povo Apinajé - parte 2
atenção básica,Povo Apinajé,Tocantinópolis

S20P207: Diario das Forcas
força,indígena,jovens,mental,saúde,suicídio

Série: Sete Pecados Capitais

S21P208: Nossa louca vontade de pecar: ensaiando conversas da presentividade
culpa,cultura,pecado,sexo,vaidade

S21P208: A tristeza pela felicidade alheia: a Inveja
culpa,inveja,pecado,religião

S21P210: Luxúria: Viagem nas asas da liberdade dos desejos mais profundos
culpa,luxúria,pecado,religião

S21P211: Comer não é pecado
culpa,gula,pecado,religião

S21P212: L'anima Dannata: A Ira no Sujeito Pós-Moderno
culpa,ira,pecado,religião

S21P213: Avareza: um pecado acumulativo que afeta as emoções
avareza,emoção,pecado,religião

S21P214: Preguiça: pecado capital ou ato de rebeldia?
culpa,pecado,preguiça,procrastinação,religião

S21P215: Tudo é Vaidade

culpa,pecado,religião,vaidade

Série: Sete Virtudes

S22P216: Castidade: a virtude da continuidade permanente
castidade,cultura,religião,virtudes

S22P217: Ela e o mar: um conto sobre a paciência
cultura,paciência,religião,virtudes

S22P218: Humildade: autoconhecimento e confiança em si mesmo
cultura,humildade,religião,virtudes

S22P219: A diligência – "Façamos, vamos amar"
cultura,diligência,religião,virtudes

S22P220: Generosidade: uma perspectiva judaica
cultura,generosidade,religião,virtudes

S22P221: Caridade: Amor em sua forma mais nobre
caridade,cultura,religião,virtudes

S22P222: A temperança e a escravidão da vontade
cultura,religião,temperança,virtudes

S22P223: O tempero das virtudes
cultura,religião,virtudes