UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

MARLON PEREIRA FARIAS

UM SISTEMA DE RECOMENDAÇÃO DE LINKS PARA O FOMENTO DE DISCUSSÕES EM FÓRUNS DE UM AMBIENTE VIRTUAL DE APRENDIZAGEM

SÃO LUÍS

2015

MARLON PEREIRA FARIAS

UM SISTEMA DE RECOMENDAÇÃO DE LINKS PARA PROMOVER DISCUSSÕES EM FÓRUNS DE UM AMBIENTE VIRTUAL DE APRENDIZAGEM

Dissertação apresentada ao Mestrado Profissional de Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão, como parte dos requisitos para a obtenção do título de Mestre em Engenharia da Computação e Sistemas.

Orientador: Prof. Dr. Luís Carlos Consta Fonseca

SÃO LUÍS

2015

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus, por estar sempre presente em minha vida, dando-me forças para superar todas as dificuldades.

Ao meu orientador, o professor Dr. Luís Carlos Costa Fonseca, pela paciência, dedicação e confiança creditados a mim durante todo o período de desenvolvimento deste trabalho.

A minha noiva, Stéfany Daniela, pelo amor, incentivo e compreensão durante todo o período do mestrado.

Aos meus pais, pelo afeto e assistência ao longo de uma vida, fazendo com que nunca desistisse dos meus sonhos e objetivos.

Aos docentes e discentes da segunda turma do Mestrado de Engenharia da Computação e Sistemas da UEMA.

A todos que, direta e indiretamente, contribuíram para a realização deste trabalho.

“O coração do inteligente adquire o conhecimento, e o ouvido dos sábios busca a sabedoria”.

Provérbios 18:15

RESUMO

Os fóruns de discussão apresentam-se como algumas das ferramentas de interação mais utilizadas nos ambientes virtuais de aprendizagem. Eles são objeto de estudo de várias pesquisas em informática na educação, tanto no que se refere a sua melhor utilização, como na avaliação de seus registros. Neste trabalho, é proposta uma ferramenta que identifica palavras-chave relevantes através de técnicas de mineração textual, e, posteriormente, às submete a um motor de busca da Internet para recuperar documentos correlacionados aos assuntos do fórum. Pretende-se, com isso, fomentar os debates através da apresentação de referencial textual extraído da rede mundial de computadores, e, dessa forma, propor uma recomendação de informações baseada em conteúdo.

Palavras-chave: Recomendação de informação. Mineração de texto. tf*pdf.

ABSTRACT

Discussion forums present themselves as one of the most popular interaction tools in learning management systems. Many researchers in computer science education study them, both concerning their best use, such as assessment of their records. This paper proposes a tool that identifies relevant keywords via text mining techniques, and then submit it to a search engine on the Internet to retrieve documents related to the matters of the forum. It is intended with this, foment debate by presenting extracted textual reference from the World Wide Web, and thus propose a content-based recommendation of information.

Keywords: Recommendation systems. Text mining. tf*pdf.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

AIM – *Articulated Instructional Media* (O Projeto Mídia de Instrução Articulada)

API – *Application Programming Interface* (Interface de Programação de Aplicações)

AVA – Ambientes Virtuais de Aprendizagem

Cier. Centro Internacional de Estudos Regulares

Consun – Conselho Universitário

e-Tec – Escola Técnica Aberta

EaD – Educação a Distância

HTTP – *Hypertext Transfer Protocol* (Protocolo de Transferência de Hipertexto)

HTTPS – *Hypertext Transfer Protocol Secure* (Protocolo Seguro de Transferência de Hipertexto)

Ibam – Instituto Brasileiro de Administração Municipal

IDF – *Inverse Document Frequence* (frequência inversa do documento)

IUB – Instituto Universal Brasileiro

JSON – JavaScript Object Notation (Notação de Objetos JavaScript)

MEB – Movimento de Educação de Base

PDF – *Ponderal Document Frequence* (Frequência do Documento Ponderada)

PLN – Processamento de Linguagem Natural

RI – Recuperação de Informação

SETEC – Secretaria de Educação Profissional e Tecnológica

TI – Tecnologia de Informação

TIC – Tecnologias de Informação e Comunicação

TF – *Term frequence* (Frequência do Termo)

UEMA – Universidade Estadual do Maranhão

UemaNet – Núcleo de Tecnologias da Educação da Universidade Estadual do Maranhão

UNED – Universidade Nacional de Educação a Distância

URI – *Uniform Resource Identifier*

XML – Extensible Markup Language (Linguagem de marcação extensível)

WWW – *World Wide Web*

SUMÁRIO

INTRODUÇÃO

Qual o melhor caminho para se chegar ao centro na hora do rush? Onde se pode encontrar um bom restaurante nas proximidades? Qual a melhor época para viajar para o Caribe? Qual a melhor empresa para investir em ações? Qual computador comprar em atendimento às necessidades específicas de um determinado usuário?

Tomar decisões é algo corriqueiro na vida das pessoas. Para isso, faz-se uso de diversas técnicas que apoiam esse processo, como por exemplo: pedir ajuda a um amigo, fazer uma consulta na Internet, contratar um especialista da área, entre outros. Nessas situações, o que se deseja é um meio que possa auxiliar a realizar boas decisões, ou seja, uma recomendação.

O ser humano está acostumado a pedir sugestões. Essa atividade é tão comum e frequente que se viu a possibilidade de realizá-la de forma automática, através de sistemas computacionais. Empresas como Amazon, NetFlix, Ebay, Youtube, Google, Yahoo, entre outras, destacaram-se ao fazer uso desses sistemas e investir no desenvolvimento de novos métodos relacionados com essa tecnologia.

Ricci et al. (2011) apresentam os Sistemas de Recomendação como ferramentas de software e técnicas que fornecem indicações de itens que sejam úteis para o usuário. Essas sugestões referem-se a vários processos de tomada de decisão, tais como: os itens a comprar, qual música ouvir, ou quais notícias online ler. O “item” é a terminologia geral utilizada para referenciar o que o sistema recomenda aos usuários.

A utilização de Sistemas de Recomendação tem crescido bastante desde o seu surgimento, na década de 1990. Desde então, muito se tem feito em termos de criação de novos métodos e abordagens que permitam a utilização dessa tecnologia nas mais diversas áreas. Os sistemas de comércio eletrônico destacam-se entre os demais como os principais utilizadores dessa tecnologia, de tal forma que ela já é parte fundamental da maior parte deles.

A educação a distância (EAD) passou por diversas mudanças nos últimos vinte anos, em razão do advento da Internet da década de 1990. O uso das Tecnologias de Informação e Comunicação (TICs), principalmente as relacionadas à Internet, possibilitou que os participantes (professores e alunos) do processo de ensino e aprendizagem pudessem interagir.

Piva et al. (2011) afirmam que essa evolução provou uma mudança paradigmática no sentido de que a individualização cedeu lugar à colaboração. E acrescenta que a

aprendizagem independente passou a ser sustentada por experiências colaborativas entre alunos e professores, e alunos entre si. Os sistemas de gerenciamento de cursos online reforçaram a socialização e passaram a ser desenhados de maneira a permitir vários tipos de interação, proporcionando meios para estimular o envolvimento e a comunicação entre os participantes, por intermédio de ferramentas ora síncronas (em tempo real), ora assíncronas (remota).

O Ambiente Virtual de Aprendizagem (AVA) é um espaço online destinado a organizar e coordenar as atividades de ensino, ou seja, são as salas de aula online. Esse ambiente é composto por diversas ferramentas, como os chats, as videoconferências, os fóruns, entre outros. Esses instrumentos auxiliam o processo de interação e buscam tornar o aprendizado mais atrativo para o aluno. Uma das principais ferramentas empregadas na construção do conhecimento nos AVA são os fóruns. Eles são instrumentos de discussão e troca de ideias, que auxiliam na construção coletiva do conhecimento e na integração entre alunos e professores. Trata-se de um espaço interativo assíncrono para a troca de mensagens, permitindo a todos os participantes trocarem ideias e conhecimentos. Os fóruns de discursão são uma das principais ferramentas voltadas para a construção de conhecimento colaborativo nos AVA. Entretanto, muitas vezes, os alunos prendem-se à aula do professor ou a uma vídeoaula, não buscando novas fontes de conhecimento, e, logo, tornando as discussões pouco proveitosas.

Fernadez (2003, apud PIVA, 2011) tece uma crítica aos AVA, afirmando que eles não são capazes de, adequadamente, manter processos de aprendizado que permitam, além de um retorno automático aos alunos, um direcionamento do processo. Diante do exposto, surge uma pergunta: Como oferecer um meio por meio do qual os participantes dos fóruns possam receber, de forma dinâmica, sugestões de leitura de acordo com o contexto da discussão, oferecendo materiais além dos pré-cadastrados na sala de aula virtual?

O presente trabalho busca por alternativas para oferecer novas fontes de conhecimento, a fim de tornar a discussão mais proveitosa através da recomendação de links.

I) Motivação

O surgimento e a expansão da Internet na década de 1990 foram responsáveis por profundas mudanças na EAD. O uso das TICs trouxe um novo formato de interação entre professores e alunos.

Uma das principais ferramentas dos AVA, voltadas à construção do conhecimento de forma colaborativa, são os fóruns de discussão, que consistem em espaços de discussões e troca de ideias em torno de temas propostos por seus participantes. Esse instrumento permite que cada participante submeta sua colaboração referente ao tema proposto, buscando, assim, o entendimento mútuo. Segundo Silva (2006 apud OKADA, 2006), o fórum é uma ferramenta assíncrona que representa um espaço para debates no qual pode ocorrer o entrelaçamento de muitas vozes para a construção e desconstrução de pensamentos, para questionar e responder dúvidas, trilhando novos caminhos para a aprendizagem.

Sobre a importância desses debates, Kenski (2002) tece o seguinte comentário:

> Interagir com o conhecimento e com as pessoas para aprender é fundamental. Para a transformação de um determinado grupo de informações em conhecimentos é preciso que estes sejam trabalhados, discutidos, comunicados. As trocas entre colegas, os múltiplos posicionamentos diante das informações disponíveis, os debates e as análises críticas auxiliam a sua compreensão e elaboração cognitiva. As múltiplas interações e trocas comunicativas entre parceiros do ato de aprender possibilitam que estes conhecimentos sejam permanentemente reconstruídos e reelaborados. (KENSKI, 2002, p. 258)

Com relação à participação e envolvimento nos fóruns de discursão, Oliveira (2005) salienta que a participação nesse espaço demanda preparo, geralmente provido por leituras adequadas, pesquisas, resgates ao *background* próprio a cada participante, entre outras formas de busca. Trata-se de organizar o pensamento, enriquecendo-o com pertinentes referências, permitindo o uso do espaço de discussões e reflexões proporcionado pelo fórum para gerar colaborações, para agregar ideias. É a chance de valorizar o conhecimento abalizado, com espaço para opiniões pessoais – e a discussão das mesmas – sem que essa iniciativa represente uma apologia ao "achismo" e ao acúmulo de debates improfícuos, destituídos de solidez teórica. O tempo comunicacional, assíncrono, favorece semelhante postura, em um espaço potencialmente livre de conflitos.

Partindo desse problema, este trabalho propõe a utilização de um sistema com o objetivo de fomentar discussões em fóruns, o que se consolida através da recomendação automática de links para estimular os alunos a buscarem leituras diferentes daquelas já trabalhadas dentro da sala virtual.

II) Objetivos

O fórum pode ser visto como um elemento assíncrono de envio de mensagens em rede, as quais são destinadas, na maioria das vezes, a um grupo de pessoas habilitadas ao

acesso das mesmas. Os fóruns de discussão representam um espaço para a construção colaborativa do conhecimento, onde as informações são trabalhadas e discutidas, para que, através dos múltiplos posicionamentos, possam ser transformadas em conhecimento.

Para que as discussões possam fluir de maneira natural estre os participantes, os professores (elaboradores do fórum) geralmente optam por não restringirem temas muitos específicos, de modo a evitar que os alunos, simplesmente, busquem por um conceito, o copiem e o colem no fórum. Oposto a isso, as temáticas dos fóruns geralmente permitem que a discussão possa tomar diferentes caminhos, oferecendo aos seus participantes possibilidades de seguirem por diferentes abordagens, sem se desviarem da temática em discussão. Nesse processo, cabe ao professor e ao tutor o papel de intermediar o processo de discussão, de modo a manter o foco na temática e estimular o debate em questão.

Devido à possibilidade de a temática do fórum fluir por diferentes vertentes, é necessário que haja uma maneira de se analisar o conteúdo em debate para que, então, possa haver a sugestão de links para leitura. Dessa forma, este trabalho tem como objetivos:

1. Desenvolver um analisador de conteúdo a fim de identificar, em meio ao conteúdo dos fóruns, quais tópicos são mais relevantes (*hot topics*);
2. Submeter o assunto a um motor de busca, de modo a capturar *links* correlacionados aos tópicos para apresentá-los aos usuários. No exemplo aqui proposto, foi utilizado a API do motor de busca Bing;
3. Investigar a pertinência e a qualidade dos tópicos extraídos; e
4. Investigar a pertinência e a qualidade dos links recuperados.

Feita a enumeração dos objetivos, prossegue-se à apresentação da organização do trabalho.

III) Apresentação do Trabalho

Este trabalho está dividido em seis capítulos, dispostos da seguinte forma: o primeiro capítulo apresenta os conceitos e o histórico da EAD; o segundo capítulo expõe os tópicos relacionados às tecnologias de Recuperação de Informação (RI), destacando o processo de ponderação de termos; no terceiro capitulo é apresentada a tecnologia de recomendação de informação e as suas principais técnicas; o quarto capítulo apresenta a tecnologia de mineração de texto; o quinto capítulo delineia a descrição do sistema implementado no decorrer desta pesquisa; e o sexto capítulo exibe a pesquisa realizada, assim

como os seus resultados. E, por fim, nas considerações finais, são apresentadas as conclusões deste trabalho.

CAPÍTULO I

A EDUCAÇÃO A DISTÂNCIA E OS AMBIENTE VIRTUAIS DE APRENDIZAGEM

Apesar do grande destaque que a EAD vem obtendo nas últimas duas décadas, essa modalidade de ensino não é tão nova como muitos acreditam. Ela apenas ganhou maior destaque devido às novas possibilidades oferecidas pelas TICs, principalmente as relacionadas à Internet.

Faria e Salvadori (2010) destacam a adaptabilidade como um dos principais fatores para o crescente interesse por essa modalidade de ensino, afirmando que:

> EaD é uma modalidade de ensino que cada vez mais está se destacando no cenário atual, principalmente porque se adapta à diferentes realidades dos alunos que procuram formação mediante este meio. A EaD não se trata de uma forma facilitada de conseguir títulos, muito menos de formação de baixa qualidade. Trata-se de um sistema que atende as necessidades de um público específico e está atingindo cada vez mais segmentos. (FARIA; SALVADORI, 2010, p. 16)

Mas, o que é EAD? Ainda neste capítulo, há uma seção destinada exclusivamente à definição desse termo. Entretanto, para que possamos entender a sua origem, apresentaremos um conceito inicial, apresentado por Maia e Mattar (2007, p. 6): "a EAD é uma modalidade de educação em que professores e alunos estão separados, é planejada por instituições e utiliza diversas tecnologias de comunicação".

Para que possamos entender melhor o contexto que levou o seu grande crescimento, será apresentado um pouco do histórico da EAD, bem como sua evolução e as transformações sofridas no decorrer do tempo.

1.1 Histórico

A EAD é a modalidade de educação em que professores e alunos estão separados fisicamente, no espaço e/ou no tempo. Maia e Mattar (2007) afirmam que ela é tão velha quanto a origem da escrita. Antes de existir a escrita, a comunicação dava-se através da linguagem oral, sendo assim, o processo de comunicação era necessariamente presencial. Para que uma informação pudesse ser transmitida, era necessário que o emissor e o receptor estivessem presentes no mesmo local e ao mesmo tempo.

A partir da invenção da escrita, a comunicação libertou-se no tempo e no espaço. Sendo assim, "não é mais necessário que as pessoas estejam presentes, no mesmo momento e local, para que haja comunicação. Em uma sociedade primitiva, ao contrário, não ocorre comunicação sem que a pessoa com quem desejamos nos comunicar esteja presente". (MAIA; MATTAR, 2007, p. 21)

Muitos autores consideram as mensagens escritas utilizadas para a difusão do Cristianismo como a primeira de iniciativa educacional, tendo como destaque as epístolas de São Paulo, escritas com o objetivo de ensinar às comunidades cristãs da Ásia Menor como viverem como cristãs em um ambiente desfavorável. Segundo Peters (2004, p. 29), "através das tecnologias da escrita e dos meios de transporte a fim de fazer seu trabalho missionário sem ser forçado a viajar o apóstolo Paulo conseguia substituir o ensino tradicional face a face por pregação e ensino assíncrono e mediados".

Vale ressaltar que a motivação que levou Paulo a fazer uso desse modelo de ensino foi a sua prisão em Roma. Muitas das suas epístolas foram escritas no período em que ele se encontrava em reclusão (incialmente em regime domiciliar, segundo Atos 28:30-31), quando, mesmo impedido de levar a mensagem de maneira tradicional, utilizou-se da escrita de cartas para quebrar as barreiras e ofertar o conhecimento, fazendo o uso das tecnologias existentes até então.

> Alguns autores consideram as cartas de Platão e as Epístolas de São Paulo exemplos iniciais e isolados de exercícios de educação a distância. Outros defendem que o ensino a distância tronou-se possível apenas com a invenção da imprensa, no século XV. A escrita, inicialmente, possibilitou que pessoas separadas geograficamente se comunicassem e documentasse informações, obras e registros. A invenção de Gutenberg, por sua vez, facilitou esse processo, permitindo que as ideias fossem compartilhadas e transmitidas para um maior número de pessoas, o que intensificou os debates, a produção e a reprodução do conhecimento. (MAIA; MATTAR, 2007, p. 21)

A maioria dos autores costuma dividir a história da EAD em três gerações. A primeira geração destaca-se pelos cursos por correspondência; a segunda é caracterizada pelo uso das novas tecnologias, como o rádio, a televisão, e também a criação das universidades abertas; e, por fim, a terceira é a geração EAD online.

1.1.1 Primeira Geração da EAD

Os primeiros registros de cursos a distância datam da década de 1720. Segundo Nunes (2009), a primeira notícia que se tem registro da introdução desse modelo de ensino

foram as aulas de taquigrafia oferecidas por Caleb Philips, em 20 março de 1728, na Gazette de Boston, nos Estados Unidos. Nesse modelo, as aulas eram por correspondência, sendo enviadas aos alunos todas as semanas.

Em 1840, na Grã Bretanha, Isaac Ptman oferecia um curso de taquigrafia por correspondência. E, Skerry's, em 1880, ofereceu cursos preparatórios para concursos públicos. Outro indício de que a EAD estava tomando forma aconteceu nos Estados Unidos, em 1891, quando foi ofertado um curso sobre segurança nas minas, tendo como organizador Thomas J. Foster. (FARIA; SALVADORI, 2010)

Há relatos de que, em 1880, na Inglaterra, houve a tentativa de se implantar um curso por correspondência, com direito a diploma. As autoridades locais barraram a ideia, entretanto, os seus autores não desistiram e foram para os Estados Unidos, onde encontram uma oportunidade, na Universidade de Chicago, para colocarem em prática as suas ideias. Assim, em 1882, surgiu o primeiro curso universitário na modalidade EAD, em que o material didático era enviado pelo correio. Outro marco histórico na EAD também se deu nos EUA, no ano de 1906, quando a *Calvert School*, em Baltimore, tornou-se a primeira escola primária a oferecer cursos por correspondência.

Segundo Preti (1996), a crescente demanda por educação não é uma consequência somente da expansão populacional, mas, sobretudo, das lutas das classes trabalhadoras por acesso à educação, ao saber socialmente produzido. A evolução dos conhecimentos científicos e tecnológicos exigiu mudanças em nível de função e de estrutura da escola e da universidade.

Sobre os motivadores que influenciaram o surgimento da EAD, Peters (2004) destaca necessidade de ensino nas colônias britânicas e francesas:

> A educação a distância se tornou ainda mais importante para quem morava longe de seus países de origem, nas colônias. Os britânicos que serviam nas colônias do Império Britânico, por exemplo, muito frequentemente não tinham a oportunidade de cursar uma universidade tradicional e tinham que se preparar sozinhos para os exames externos da Universidade de Londres. Tinham em seu auxílio os serviços de várias faculdades por correspondência que contavam com a tecnologia do transporte marítimo a fim de fazer a entrega do material didático. Estes processos de ensino e aprendizagem em particular eram verdadeiramente assíncronos devido ao longo tempo que levava para chegar, por exemplo, aos alunos na Índia e na Austrália. Estas atividades representam outra raiz da educação a distância e da educação superior aberta moderna. O mesmo pode ser dito sobre alunos a distância nas colônias francesa matriculados na “Escole Universelle” de Paris. (PETERS, 2004, p. 31)

A primeira geração da EAD aconteceu no período entre 1728 até meados de 1970, tendo como principal característica o estudo por correspondência. Nesse período, as

possibilidades de interação entre aluno e instituição produtora eram bastante restritas, e, geralmente, limitavam-se apenas aos momentos de exames. Didaticamente, os alunos recebiam o material impresso para estudos, o que era acompanhado por exercícios de fixação.

Apesar das divergências dos autores quanto à primeira experiência a distância, deve-se deixar claro que se tratam de marcos iniciais para a expansão dessa modalidade de ensino.

1.1.2 Segunda Geração da EAD

A partir da década de 1960, a EAD assumiu novos formatos. A utilização de novas tecnologias de comunicação, como a televisão, o rádio, as fitas de áudio e vídeo e o telefone, permitiram que houvesse uma maior dinâmica nas aulas, antes restritas à leitura dos materiais.

É importante ressaltar que essas tecnologias não começaram a ser utilizadas com fins didáticos somente a partir dos anos 1960. Assim como já apresentado neste trabalho, experiências de EAD puderam ser claramente visualizadas no ensino através de cartas de São Paulo, desde meados dos anos 60 d.C.. Entretanto, foi a partir de 1728 que a EAD passou a exibir uma estrutura formal, contando com uma tecnologia que oferece maior suporte. Nesse caso, a imprensa permitiu uma comunicação em massa através da escrita.

A segunda geração tem como marco inicial o uso de outros modelos de EAD, como o rádio e a televisão, muito embora haja registros anteriores de iniciativas com esses modelos. No Brasil, por exemplo, a Rádio Sociedade do Rio de Janeiro, em 1923, transmitia programas educacionais. Porém, foi nos anos 1960 que se efetivaram as maiores experiências como esses novos modelos. (FARIA; SALVADORI, 2010)

Desde que o rádio surgiu como nova tecnologia, no início do século XX, despertou o interesse de muitos educadores dos departamentos de extensão das universidades, gerando grande entusiasmo em relação às novas possibilidades oferecidas por esse novo meio de comunicação em massa. Segundo Saettler (1990 apud MOORE; KEARSLEY, 2007, p. 32), “a primeira autorização para uma emissora educacional foi concedida pelo governo federal à Latter Day Saint’s da University of Salt Lake City em 1921”.

Faria e Salvadori (2010) afirmam que, impulsionada pelas novas tecnologias, a EAD tornou-se mais aberta em dois sentidos: oferecer maiores oportunidades de escolha temática e de tempo aos alunos; e, também, propiciar um tratamento mais personalizado, em

atendimento às necessidades individuais, demonstrando que houve a superação de um modelo industrializado de educação.

Outro marco ocorrido nesse período foi a criação da Universidade Aberta. Em 1967, o governo britânico criou um comitê para planejar uma nova e revolucionária instituição educacional. A priori, desejava-se simplesmente fazer uso das novas tecnologias de comunicação em massa (rádio e a televisão), com vistas a permitir o acesso à educação superior para a população adulta. Em novembro de 1967, autoridades do comitê do projeto visitaram Wisconsin para estudarem os métodos e realizações do projeto AIM[1], convidando Wedemeyer para se reunir em eles em Londres.

O que surgiu foi a primeira universidade nacional de educação a distancia, que se valeria de economias de escala, tendo mais alunos do que qualquer outra universidade, com um nível de financiamento elevado e empregando a gama mais completa de tecnologias de comunicação para ensinar um currículo universitário completo a qualquer adulto que que desejasse receber essa educação. (MOORE; KEARSLEY, 2007)

> AIM - O Projeto Mídia de Instrução Articulada (AIM – Articulated Instructional edia Project) – financiado pela Carnegie Corpopation de 1964 a 1968 e dirigido por Charles Wedemeyer, da University of Wisconsin em Madison – era testar a ideia de agrupar (isto é, articular) várias tecnologias de comunicação, com o propósito de oferecer um ensino de alta qualidade e custo reduzido a alunos não-universitários. As tecnologias incluíam guias de estudo impressos e orientação por correspondência, transmissão por rádio e televisão, audioteipes gravados, conferências por telefone, kits para experiência em casa e recurso de uma biblioteca local. Também articulado no programa havia o suporte e a orientação para o aluno, discussões em grupos de estudos locais e o uso de laboratórios das universidades durante o período de férias. (MOORE; KEARSLEY, 2007, p. 32)

Segundo Mugnol (2009), um dos marcos históricos da EAD foi a criação da Universidade Aberta de Londres, em 1970, a *Open University*, que contribuiu decisivamente para o desenvolvimento de métodos e técnicas que serviram para caracterizar os diferentes modelos de EAD existentes. Além disso, colaborou, também, para o desenvolvimento de tecnologias que deram mais solidez aos processos educacionais a distância e para a utilização massiva da mídia. Seguindo o exemplo da Inglaterra, outros países criaram instituições para desenvolver projetos formais de EAD. Assim, no ano de 1972, em Madri, surgiu a Universidade Nacional de Educação a Distância (UNED), que pode ser caracterizada como uma das iniciativas de maior sucesso e que serviu de modelo para outros países.

O uso dos meios eletrônicos de comunicação em massa, como o rádio, a televisão e a criação das universidades abertas, foi o ponto marcante dessa fase, trazendo novas possibilidades de ensino e, sobretudo, estruturação e organização para a EAD.

1.1.3 Terceira Geração da EAD

A terceira geração foi marcada pelo desenvolvimento das TICs. O advento da Internet da década de 1990 permitiu uma mudança muito grande na EAD. O uso das TICs, principalmente as relacionadas à Internet, possibilitou que os participantes (professores e alunos) do processo de ensino e aprendizagem pudessem interagir.

Criada em 1989 por Tim Berners-Lee, a *World Wide Web* (WWW ou simplesmente Web) consiste em um sistema de documentos em hipermídia que são interligados e executados na Internet. Para visualizar a informação, o usuário faz uso de um programa chamado navegador, que descarrega essas informações e as mostra na tela do usuário. Seu grande sucesso está relacionado ao fato de possuir uma interface de usuário padrão, não dependendo do ambiente computacional utilizado para ser executada, e permitir que qualquer usuário crie seus próprios documentos.

"Uma terceira geração introduziu a utilização do videotexto, do microcomputador, da tecnologia de multimídia, do hipertexto e de redes de computadores, caracterizando a EAD online. Além disso, em relação à geração anterior, não há mais uma diversidade de mídias que se relacionam, mas uma verdadeira integração de todas delas, que convergem para as tecnologias de multimídia e do computador". (MAIA; MATTAR, 2007, p. 22)

Para Faria e Salvadori (2010), trata-se de uma nova tendência na EAD, caracterizada, sobretudo, pela flexibilidade proporcionada pela integração de várias tecnologias, como, por exemplo, a telemática (informática com telecomunicação). A aplicação das novas tecnologias da informação na educação gera condições para que o aprendizado seja cada vez mais interativo e autônomo. O estudante determina seu tempo, seu ritmo e tem acesso, em qualquer lugar e em todo tempo, aos recursos necessários através do computador conectado à Internet.

1.1.4 Evolução da EAD no Mundo

Sobre a evolução histórica da EAD no mundo, Alves (2011) destaca alguns marcos importantes:

- **1728** – marco inicial da EAD: é anunciado um curso pela Gazeta de Boston, na edição de 20 de março, em que o professor Caleb Philipps, de *Short Hand*, oferecia material para ensino e tutoria por correspondência.

- **1829** – na Suécia é inaugurado o Instituto Líber Hermondes, que possibilitou a mais de cento e cinquenta mil pessoas realizarem cursos através da EAD;
- **1840** – na Faculdade Sir Isaac Pitman, no Reino Unido, é inaugurada a primeira escola por correspondência na Europa;
- **1856** – em Berlim, a Sociedade de Línguas Modernas patrocina os professores Charles Tous-saine e Gustav Laugenschied para ensinarem Francês por correspondência;
- **1892** – no Departamento de Extensão da Universidade de Chicago, nos Estados Unidos da América, é criada a Divisão de Ensino por Correspondência para preparação de docentes;
- **1922** – iniciam-se cursos por correspondência na União Soviética;
- **1935** – o Japanese National Public Broadcasting Service inicia seus programas escolares pelo rádio, como complemento e enriquecimento da escola oficial;
- **1947** – inicia-se a transmissão das aulas de quase todas as matérias literárias da Faculdade de Letras e Ciências Humanas de Paris, na França, por meio da Rádio Sorbonne;
- **1948** – na Noruega, é criada a primeira legislação para escolas por correspondência;
- **1951** – nasce a Universidade de Sudáfrica, atualmente a única a distância da África, que se dedica exclusivamente a desenvolver cursos nesta modalidade;
- **1956** – a Chicago TV College, nos Estados Unidos, inicia a transmissão de programas educativos pela televisão, cuja influência pode se notar rapidamente em outras universidades do país, que não tardaram em criar unidades de ensino a distância baseadas fundamentalmente na televisão;
- **1960** – na Argentina, nasce a Tele Escola Pri-mária do Ministério da Cultura e Educação, integrando os materiais impressos à televisão e à tutoria;
- **1968** – é criada a Universidade do Pacífico Sul, uma universidade regional pertencente a doze países-ilhas da Oceania;
- **1969** – no Reino Unido, é criada a Fundação da Universidade Aberta;
- **1971** – a Universidade Aberta Britânica é fundada;
- **1972** – na Espanha, é fundada a Universidade Nacional de Educação a Distância;
- **1977** – na Venezuela, é criada a Fundação da Universidade Nacional Aberta;
- **1978** – na Costa Rica, é fundada a Universidade Estadual a Distância;
- **1984** – na Holanda, é implantada a Universidade Aberta;
- **1985** – é criada a Fundação da Associação Europeia das Escolas por Correspondência;

- **1985** – na Índia, é realizada a implantação da Universidade Nacional Aberta Indira Gandhi;
- **1987** – é divulgada a resolução do Parlamento Europeu sobre Universidades Abertas na Comunidade Europeia;
- **1987** – é criada a Fundação da Associação Europeia de Universidades de Ensino a Distância;
- **1988** – em Portugal, é criada a Fundação da Universidade Aberta;
- **1990** – é implantada a rede Europeia de Educação a Distância, baseada na declaração de Budapeste, no relatório da Comissão sobre Educação Aberta e a Distância na Comunidade Europeia.

1.1.5 Evolução da EAD no Brasil

Maia e Mattar (2007) apresentam um pouco da evolução da EAD no Brasil, destacando alguns fatos e datas:

- 1907 – Escolas internacionais e cursos por correspondência. Considera-se marco histórico a implantação das "Escolas Internacionais", que representam organizações norte-americanas. Eram instituições privadas que ofereciam cursos pagos por correspondência em jornais, sendo que, inicialmente, as aulas eram em espanhol.
- 1923 – Rádio-Escola. Um grupo liderado por Henrique Morize e Roquette-Pinto criou a Rádio Sociedade do Rio de Janeiro, que oferecia cursos de português, francês, silvicultura, literatura francesa, esperanto, radiotelegrafia e telefonia, dando início à educação pelo rádio.
- 1939 – Rádio Monitor. Os primeiros institutos a oferecerem cursos profissionalizantes a distancia foram a Rádio Técnico Monitor e o Instituto Universal Brasileiro, em 1941.
- 1941 – IUB. O Instituto Universal Brasileiro foi fundado por um ex-sócio do Instituto Monitor e já formou mais de quatro milhões de pessoas. Oferece cursos profissionalizantes, como auxiliar de contabilidade, desenho artístico e publicitário, fotografia, inglês, violão, e outros.
- 1943 – A Voz da Profecia. Iniciada nos Estados Unidos, em 1929, com a transmissão de séries bíblicas por rádio. Em 1943, passaram a ser gravados discos e transmitidos

programas por rádio, em português. Assim, foi ao ar o primeiro programa religioso apresentado no Brasil pela rádio.

- 1947 – Senac, Sesc e Universidade do Ar. O Sesc, Senac e emissoras associadas fundam a Universidade do Ar, com objetivo de oferecer cursos comerciais radiofônicos. Os alunos estudavam nas apostilas e corrigiam exercícios com o auxílio dos monitores.
- 1961 – MEB. O Movimento de Educação de Base foi criado pela Diocese de Natal. Foram escolas radiofônicas voltadas para a democratização do acesso à educação, promovendo o letramento de jovens e adultos.
- 1962 – Ocidental School. Fundada em São Paulo, a Ocidental School, de origem americana, focava-se no campo da eletrônica.
- 1967 – Ibam. Instituto Brasileiro de Administração Municipal iniciou suas atividades de EAD utilizando a metodologia de ensino por correspondência.
- 1967 – Padre Landell. A fundação Padra Landell de Moura criou seu núcleo de EAD com metodologia de ensino por correspondência e via rádio.
- 1967 – Projeto Saci. Concebido experimentalmente em 1967, por iniciativa do Instituto Nacional de Pesquisas Espaciais, o Projeto Saci (Satélite Avançado de Comunicação Interdisciplinares) tinha como objetivo criar um sistema nacional de telecomunicações por meio do uso de satélite.
- 1970 – Projeto Minerva. O projeto Minerva foi um feito por meio de um convênio entre o Ministério da Educação, a Fundação Padra Landell de Moura e a Fundação Padra Anchieta, cuja meta era utilização do rádio para a educação e a inclusão social de adultos.
- 1977 – Telecurso. Cursos supletivos a distância começaram a ser oferecidos por fundações privadas e organizações não governamentais a partir das décadas de 1970 e 1980, utilizando tecnologia de teleducação, satélite e materiais impressos. Na década de 1970, a Fundação Roberto Marinho lançou o programa de educação supletiva a distância para primeiro e segundo graus. Hoje denominado Telecurso 2000, utiliza livros, vídeos e transmissão por TV.
- 1981 – Cier. O centro Internacional de Estudos Regulares do Colégio Anglo-Americano, fundado em 1981, oferece ensinos fundamental e médio a distancia. O objetivo do Cier é permitir que crianças cujas famílias se mudam temporariamente para o exterior continuem a estudar pelo sistema educacional brasileiro.

- 1991 – Salto para o Futuro. O programa Jornal da Educação – Edição do Professor, concebido e produzido pela Fundação Roquette-Pinto, teve início em 1991. Em 1995, com o nome de Salto para o Futuro, foi incorporado à TV Escola (canal educativo da Secretaria de Educação a Distância do Ministério da Educação), tornando-se um marco na EAD nacional.

1.2 Conceitos

Cada vez mais, a EAD destaca-se no cenário atual, principalmente pela sua capacidade de se adaptar às diferentes realidades vivenciadas pelos alunos que procuram por uma formação. Não se trata de uma forma facilitada de conseguir uma formação, tampouco uma formação de baixa qualidade, mas sim consiste em uma modalidade que atende às necessidades de um público que cresce cada vez mais.

A EAD mostra-se um instrumento fundamental para promover oportunidades a pessoas que não teriam a possibilidade de obter uma formação através do ensino tradicional, pois, muitas vezes, o tempo disponível por esses alunos não se adequa ao modelo de ensino tradicional, e, por isso, eles acabam buscando um ensino que se adeque a sua realidade.

Mas, o que é EAD? Diferentes autores destacam distintos aspectos dessa modalidade de ensino. Segundo Holmberg (1985 apud ALVES, 2011), a expressão "educação à distância" cobre as distintas formas de estudo, em todos os níveis, que não se encontram sob a contínua e imediata supervisão dos tutores, presentes com seus alunos na sala de aula, mas que, não obstante, beneficiam-se do planejamento, orientação e acompanhamento de uma organização tutorial.

Dohmem (1967 apud ALVES, 2011) apresenta um conceito de EAD no qual destaca a autonomia do aluno e afirma que a esse formado de educação é uma forma sistematicamente organizada de autoestudo, em que o aluno instrui-se a partir do material de estudo que lhe é apresentado, e o acompanhamento e a supervisão do sucesso do estudante são levados a cabo por um grupo de professores. Isso é possível através da aplicação de meios de comunicação capazes de vencer longas distâncias.

Keegan (1991 apud ALVES, 2011) destaca a separação física entre professor e aluno, mas ressalta a importância do diálogo. O autor define a EAD como a separação física entre professor e aluno, aspecto que a distingue do ensino presencial. Trata-se de uma comunicação de mão dupla, em que o estudante beneficia-se de um diálogo e da possibilidade de iniciativas

de dupla via, com probabilidade de encontros ocasionais com propósitos didáticos e de socialização.

Chaves (1999 apud ALVES, 2011) salienta que a separação física é suprida pelas tecnologias de telecomunicação. A separação entre professor e aluno é provida através de tecnologias de telecomunicação e de transmissão de dados, voz e imagens (incluindo dinâmicas, isto é, televisão ou vídeo). Não é preciso ressaltar que todas essas tecnologias, hoje, convergem para o computador.

De acordo com Guarezi (2009, p. 20), "os conceitos de EAD mantêm em comum a separação física entre o professor e o aluno, e a existência de tecnologias para mediar a comunicação e o processo de ensino aprendizagem". A evolução do conceito se dá em relação aos processos de comunicação, pois a EAD, cada vez mais, passa a possuir maiores possibilidades tecnológicas para efetivar a interação entre os pares para aprendizagem.

De forma geral, os conceitos de EAD apontam três características importantes:

1) A separação entre aluno e professor (podendo ser em relação a tempo e espaço);
2) O planejamento; e
3) O uso de TICs.

Sempre que se pensa, logo se imagina a separação entre aluno e professor, ou seja, eles não estão presentes fisicamente em um mesmo local. Para a EAD, a presença física em um mesmo local não é precedente para que se promova a educação. O aprendizado não pode estar restrito à sala de aula. Além da separação física, essa modalidade de educação também está associada à separação temporal. Apesar de haver diversas atividades síncronas, como encontros presenciais, chats, videoconferências, e outras, a maior parte delas são assíncronas, permitindo que os alunos estudem quando lhes for mais conveniente. Outro aspecto importante que se deve destacar é o planejamento das atividades. A EAD não é um autoestudo espontâneo, haja vista que exige um apoio e deve ser planejada por uma instituição de ensino.

O uso de tecnologias de comunicação é o meio que permite superar a barreira da distância em termos de tempo e espaço. No decorrer da evolução da EAD, as TICs desempenharam um papel fundamental para o crescimento e expansão dessa modalidade de ensino, fazendo assumir diferentes formatos no decorrer do tempo, podendo alcançar um público cada vez maior. De modo geral, pode-se afirmar que a EAD aplica as tecnologias disponíveis para fazer acontecer os processos de ensino e aprendizagem, superando as barreiras do espaço e do tempo. "Dentre as principais características da EaD, deve-se

fortalecer aquelas ligadas a autonomia do estudante, a comunicação e o processo tecnológico, e assim é possível construir um conceito mais completo". (GUAREZI, 2009, p. 20)

Segundo Silva e Borba (2010), os cursos de EAD passaram a representar o novo contexto de aprendizagem no mundo moderno, o que ultrapassa o âmbito da sala de aula, ampliando-se e se adaptando às atuais condições dos sujeitos, oferecendo, através das novas tecnologias, a possibilidade de desenvolvimento autônomo de sua aprendizagem e formação profissional.

1.3 Regulamentação da EAD no Brasil

As bases legais para a modalidade de EAD no Brasil são estabelecidas pela Lei nº 9.394, de 20 de dezembro de 1996 - Lei de Diretrizes e Bases da Educação Nacional - e regulamentada pelo Decreto nº 5.622, publicado no Diário Oficial da União em 20 de dezembro de 2005, que apresenta a seguinte definição:

> Art.1º Para os fins deste Decreto, caracteriza-se a Educação a Distância como modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorre com a utilização de meios e tecnologias de informação e comunicação, com estudantes e professores desenvolvendo atividades educativas em lugares ou tempos diversos. (BRASIL, 1996, s/p)

Mais a diante, esse mesmo artigo complementa e ressalta a obrigatoriedade dos momentos presenciais, como segue: ***§ 1º A Educação a Distância organiza-se segundo metodologia, gestão e avaliação peculiares, para as quais deverá estar prevista a obrigatoriedade de momentos presenciais para:***

1) avaliações de estudantes;
2) estágios obrigatórios, quando previstos na legislação pertinente;
3) defesa de trabalhos de conclusão de curso, quando previstos na legislação pertinente e
4) atividades relacionadas a laboratórios de ensino, quando for o caso.

Essa normatização deixa claro que a EAD não é uma modalidade de ensino de qualidade inferior às tradicionais. Não há dúvida de que existe preconceito, porém, a EAD não carece de aparato legal e, por vezes, falta conhecimento dos próprios profissionais que atuam na área sobre a legislação que ampara o trabalho que desempenham. O aspecto legal é um ponto fundamental para derrubar preconceitos com relação a essa modalidade de

educação, de modo a deixar cada vez mais claro sua seriedade, e, ainda, realçar o fato de que se pode aprender por meio ela tanto quanto através de um modelo tradicional. (FARIA; SALVADORI, 2010)

1.4 Ambiente Virtual de Aprendizagem

A EAD passou por diversas mudanças, principalmente nos últimos vinte anos, em razão do advento da Internet da década de 1990. O uso das TICs, principalmente as relacionadas à Internet, possibilitou que os participantes (professores e alunos) dos processos de ensino e aprendizagem pudessem interagir.

O AVA é um espaço online destinado a organizar e coordenar as atividades de ensino, ou seja, são as salas de aula online. Esse ambiente é composto por diversas ferramentas, tais como os chats, as videoconferências, os fóruns, entre outros. Esses instrumentos auxiliam o processo de interação e buscam tornar o aprendizado mais atrativo para o aluno.

> A aplicação de novas tecnologias na EAD, especialmente aquelas ligadas a Internet, vem modificando o panorama dentro deste campo de tal modo que seguramente podemos falar de uma EAD antes e depois da Internet. Antes da Internet tínhamos uma EAD que utilizava apenas tecnologias de comunicação de um-para-muitos (rádio e TV) ou de um-para-um (ensino por correspondência). Via Internet temos as três possibilidades de comunicação reunidas numa só mídia: um-para-muitos, um-para-um, e sobretudo muitos-para-muitos. É esta possibilidade de interação ampla que confere à EAD via Internet um outro status e vem levando a sociedade a olhar para ela de uma maneira diferente daquela com que olha outras formas de EAD. (PIVA, 2011, p. 10)

Para Almeida (2003), AVA são sistemas computacionais disponíveis na Internet, destinados ao suporte de atividades mediadas pelas tecnologias de informação e comunicação. Eles permitem integrar múltiplas mídias, linguagens e recursos, apresentar informações de maneira organizada, desenvolver interações entre pessoas e objetos de conhecimento, elaborar e socializar produções, tendo em vista atingir determinados objetivos.

Os AVA são facilitadores da EAD por permitem interação assíncrona e síncrona entre alunos e professores tutores, o que se realiza através de ferramentas que variam de acordo com cada ambiente. Os ambientes, em sua maioria, apresentam um modelo básico, no qual as estruturas das páginas já estão definidas, e um conjunto adicional de recursos pode ser acrescentado à estrutura do curso. A criação do curso é feita através do preenchimento de formulários que geram automaticamente suas páginas e os recursos adicionais selecionados,

que, normalmente, são constituídos de ferramentas de comunicação, segurança de acesso, estatísticas de uso, acesso a banco de dados, elaboração de exercícios, etc.. (ROMANI, 2000)

Dessa forma, os AVA sanaram um dos maiores problemas da EAD tradicional, segundo Piva (2011), o chamado "isolamento" do estudante, que não contava com o apoio e o estímulo de um grupo de pessoas que estão vivenciando as mesmas experiências que ele, e apresentou a possibilidade de ajuda mútua para vencer as dificuldades.

Os recursos desses ambientes digitais de aprendizagem são basicamente os mesmos existentes na Internet (correio, fórum, bate-papo, conferência, banco de recursos, e outros), com a vantagem de propiciar a gestão da informação segundo critérios preestabelecidos de organização, definidos de acordo com as características de cada software. Possuem bancos de informações representadas em diferentes mídias (textos, imagens, vídeos, hipertextos), e interligadas com conexões constituídas de links internos ou externos ao sistema. (ALMEIDA, 2003) Os AVA buscam oferecer suporte de atividades mediadas pelas TICs. Eles oferecem suporte e gestão à sala de aula virtual, integrando diversas tecnologias.

Holmberg (1986 apud PIVA, 2011) descreve a EAD como conversação didática guiada. Para ele, estudo a distância é autoestudo, mas não deve conter apenas leitura individual, e o aluno não deve estar sozinho. Os alunos se beneficiam de ter um curso desenvolvido para eles, e também da interação com o tutor e outros elementos do suporte organizacional. É essa relação entre os alunos e o suporte organizacional que Holmberg caracteriza como conversação didática guiada, podendo ser tanto real (por meio das mídias à disposição do aluno) como simulada (estilo conversacional, por meio do estudo de textos).

Uma das principais ferramentas para a construção do conhecimento nos AVA são os fóruns, que consistem em ferramentas de discussão e troca de ideias, auxiliando na construção coletiva do conhecimento e na integração entre alunos e professores. Trata-se de um espaço interativo assíncrono voltado para a troca de mensagens, permitindo a todos os participantes trocarem ideias e conhecimentos. Os fóruns de discursão são umas das principais ferramentas na construção de conhecimento colaborativo nos AVA, visto promoverem o diálogo, o debate, a participação autônoma e interativa entre os demais participantes.

Para Oliveira (2005), o fórum pode ser visto como um elemento assíncrono de envio de mensagens em rede, destinadas, na maioria das vezes, a um grupo de pessoas habilitadas ao acesso das mesmas, cujos "direitos" são definidos por um organizador, participante ou não das interações promovidas (*designer*, em algum nível, e/ou *administrador* – um termo apropriado das definições vigentes em redes computacionais dos mais diversos tipos). Em um curso oferecido através de um AVA colaborativo, o fórum pode ser definido

como um espaço de discussões em torno de temas propostos por seus participantes. Nessa direção, o fórum parece ser o instrumento mais adequado para o aprofundamento reflexivo dos usuários do ambiente mencionado.

Kenski (2002) destaca que a interação com o conhecimento e com as pessoas é algo fundamental para o aprendizado. Para a transformação de um determinado grupo de informações em conhecimentos, é preciso que elas sejam trabalhadas, discutidas, comunicadas. As trocas entre colegas, os múltiplos posicionamentos diante das informações disponíveis, os debates e as análises críticas auxiliam a sua compreensão e elaboração cognitiva. As múltiplas interações e trocas comunicativas entre parceiros do ato de aprender possibilitam que os conhecimentos sejam permanentemente reconstruídos e reelaborados.

Para Silva e Borba (2010, p.07, grifo nosso), "o fórum possibilita uma discussão aberta e de longa duração, podendo se prolongar até a conclusão da disciplina, ou ser estipulado um prazo menor para sua contribuição. Por esse motivo, **a escolha dos temas a serem discutidos deve possibilitar uma discussão mais ampla, em que a troca de experiências entre os participantes ocorra através da soma ou contraposição de ideias apresentadas.** É importante que os alunos sejam instigados ao debate, ao questionamento das ideias do outro, e possam formular opiniões acerca do tema discutido".

CAPÍTULO II

RECUPERAÇÃO DE INFORMAÇÃO

Durante o período entre 1940 e 1950, houve um grande crescimento no volume de informações, especialmente de material bibliográfico produzido pela comunidade científica, demonstrando como a ciência poderia ser utilizada como prática e benefício para a guerra. Em consequência disso, viu-se a necessidade de obtenção de informações em um curto espaço de tempo. Mas, como encontrar uma informação específica em meio a um volume tão grande de dados? Pensando em como solucionar esse problema, pesquisadores de diversas áreas empreenderam esforços para criar técnicas de organização das informações, estratégias de busca e criação de mecanismos de busca, com o objetivo de atender com rapidez e precisão as necessidades de informação dos usuários.

A essa área de pesquisa, Mooers (1951) chamou de Recuperação de Informação (*Information Retrieval*) (RI), visto que trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, bem como qualquer sistema, técnicas ou máquinas que são empregadas para realizar essas operações. Até pouco tempo, a pesquisa de RI era vista como uma área de interesses restritos, principalmente de bibliotecários e especialistas da informação. Essa visão tendenciosa prevaleceu durante muito tempo, mas um acontecimento marcante mudou de uma vez por todas essas percepções: o surgimento da *World Wide Web*.

Criada em 1989 por Tim Berners-Lee, a *World Wide Web* consiste em um sistema de documentos em hipermídia, os quais são interligados e executados na Internet. Para visualizar a informação, o usuário faz uso de um programa chamado navegador, de modo a descarregar essas informações e mostrá-las na tela do usuário. Seu grande sucesso está relacionado ao fato de possuir uma interface de usuário padrão, não dependendo do ambiente computacional utilizado para ser executado, o que permite que qualquer usuário crie seus próprios documentos. Como resultado, segundo Yates e Ribeiro Neto (2011), a Web se tornou um repositório universal do conhecimento e da cultura humana, em que milhões de usuários criaram bilhões de documentos componentes do maior repositório de conhecimento humano na história.

> Uma consequência imediata é que encontrar informações úteis sobre a Web nem sempre é uma tarefa simples e requer geralmente colocar uma consulta a um motor de busca, ou seja, executar uma pesquisa. E a pesquisa tem tudo haver com recuperação de informação e suas tecnologias. Assim, quase que

> do dia para a noite, a recuperação de informação ganhou destaque no centro do palco junto com outras tecnologias. (YATES; RIBEIRO NETO, 2011, p. 03)

A área de RI remete à representação, armazenamento, organização e acesso a itens de informação. A representação e a organização dos itens deve prover, para o usuário, acesso fácil àquelas informações pelas quais ele se interessa. Segundo Yates e Ribeiro Neto (1999, p. 3), a "recuperação efetiva de informações relevantes é diretamente afetada pelas tarefas com as quais o usuário está envolvido *(user tasks)*, bem como pela visão lógica dos documentos adotada pelo sistema de RI". A Figura 1 ilustra a interação do usuário nas diferentes tarefas identificadas.

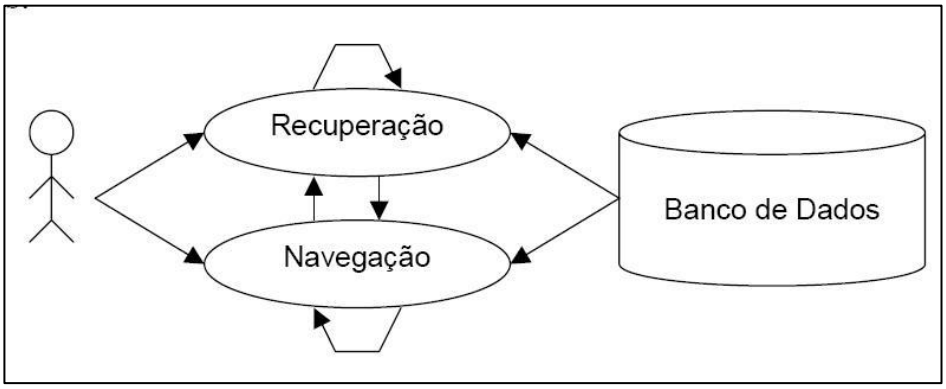


Figura 1: Interação do usuário com o sistema de recuperação através de diferentes tarefas.
Fonte: Yates; Ribeiro Neto 1999, p. 4.

2.1 A Visão Lógica de um Documento

Por questões históricas e de limitações computacionais, os documentos de uma coleção são frequentemente representados por um conjunto de termos indexadores ou por palavras-chave. Essas palavras podem ser extraídas diretamente de um documento de texto, ou especificadas por um indivíduo. Desconsiderando sua origem, quer sejam extraídas automaticamente, quer sejam geradas por um especialista, elas fornecem uma visão lógica do documento.

Os computadores modernos têm possibilitado a representação de documentos por seu conjunto completo de palavras. Nesse caso, Yates e Ribeiro Neto (1999) dizem que o sistema de recuperação utiliza uma visão lógica baseada no texto completo (uma representação). Entretanto, mesmo os computadores mais modernos, no caso de coleções muito grandes, precisam reduzir o conjunto de palavras-chave representativas. Isso pode ser

acompanhado da remoção de *stopwords* (como artigos e conetivos), o uso de *stemming* (que substituem palavras flexionadas por seus respectivos radicais), identificação de substantivos (que eliminam adjetivos, advérbios e verbos) e a compressão das estruturas lógicas internas (índices e índices invertidos, por exemplo). Essas operações, chamadas de *operações textuais*, reduzem a complexidade da representação dos documentos, além de possibilitarem a conversão de um texto completo em um conjunto de termos indexadores.

Para Yates e Ribeiro Neto (1999, p. 5), "o documento completo é claramente a mais integral visão lógica de um documento, mas seu uso implica custos computacionais elevados". Um conjunto resumido de categorias (geradas por um especialista humano) provê a visão lógica mais concisa de um documento, mas seu uso pode levar a uma recuperação de baixa qualidade. Em alguns casos, visões lógicas intermediárias (de um documento) podem ser utilizadas por um sistema de RI, como ilustra a Figura 2. Portanto, adotando algumas das representações intermediárias, o sistema de RI pode reconhecer estruturas internas normalmente encontradas em um documento (capítulos, seções, subseções, e outras).

Conforme ilustrado na Figura 2, o problema de representar logicamente um documento é contínuo. Essa representação pode variar desde o texto completo até uma especificação de alto nível, feita por um especialista humano. A escolha de uma representação mais ou menos elaborada depende da necessidade do sistema de RI a ser implementado.

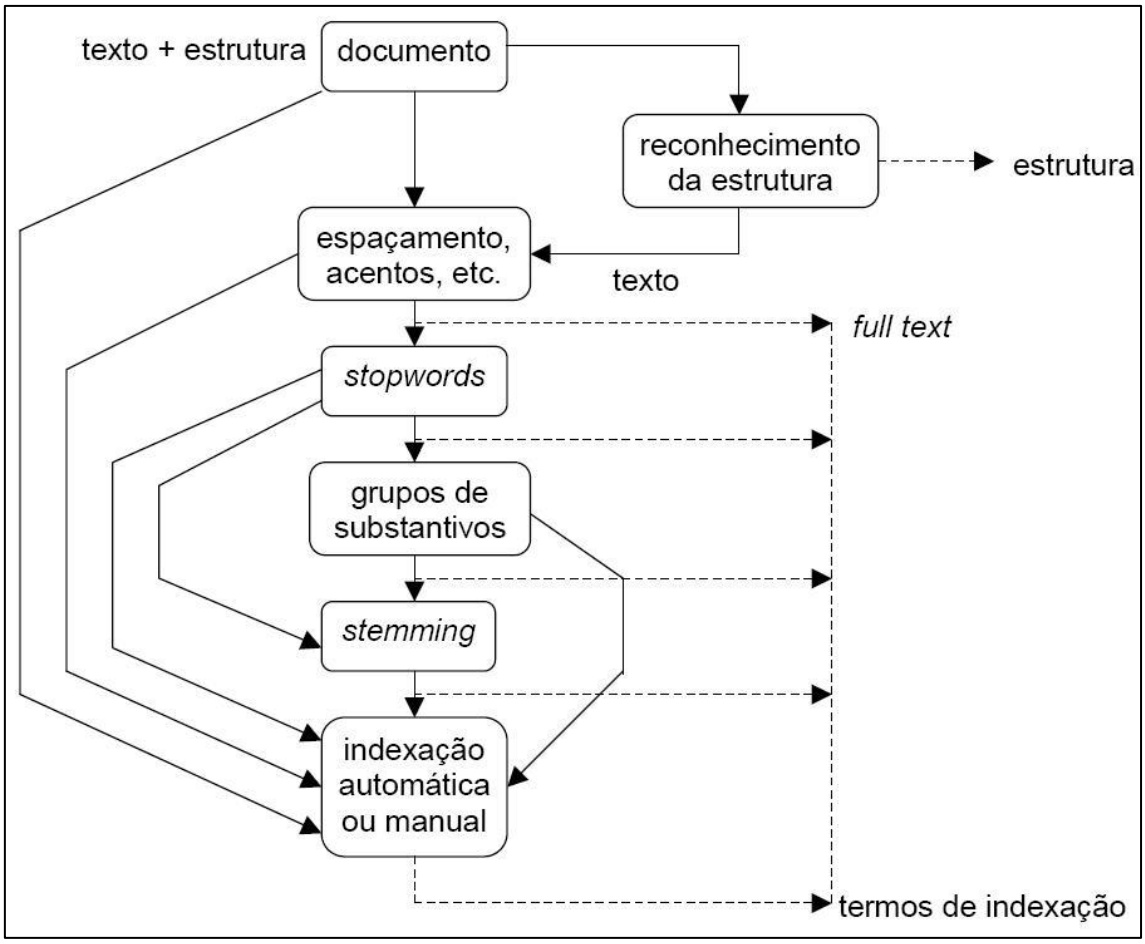


Figura 2: Visão Lógica de um Documento: do Texto Completo ao Conjunto de Termos Indexados.
Fonte: Yates; Ribeiro Neto, 1999, p. 6.

Mais a frente, serão tratadas algumas questões técnicas sobre como os documentos de uma coleção podem ser recuperados. Serão apresentados os principais modelos de RI, bem como seus princípios de funcionamento.

2.2 Term Frequency and Weighting

Segundo Salton e Buckley (1988), a principal função de um sistema de ponderação de termos (*Term Weighting*) é o aumento da eficácia na recuperação. A recuperação eficaz depende de dois fatores principais: primeiro, os itens que possam ser relevantes para as necessidades do usuário devem ser recuperados; segundo, os itens passíveis de ser estranhos devem ser rejeitados. Para tanto, Salton e Buckley (1988) apresentam duas medidas para a avaliação da capacidade de um sistema, sendo elas a precisão e o *recall*. O *recall* corresponde à proporção do número de itens relevantes recuperados para o número total

de itens relevantes na coleção. A precisão, por outro lado, compreende o número de itens relevantes recuperados e o número total de itens recuperados.

Na construção de um sistema de ponderação de termos, deve-se levar em conta ambos os requisitos. Os termos que são frequentemente mencionados em documentos, ou em trechos de documentos, parecem ser úteis. Isso sugere que o fator frequência do termo (tf) pode ser usado como parte do sistema de ponderação de termo, medindo a frequência de ocorrência dos termos nos textos de documentos ou consulta.

Definição: Para caracterizar a importância de um termo, associa-se um peso **Wi,j** >0 com cada termo **Ki** que ocorre em um documento **dj.** Se **Ki** não aparece em um documento ***dj***, então ***Wi,j*** = 0. O peso ***Wi,j*** quantifica a importância do termo índice ***Ki*** para a descrição do conteúdo do documento ***dj*.** Esses pesos são úteis para calcular uma classificação para cada documento no acervo em relação a uma determinada consulta. O peso ***Wi,j***, pode ser calculado usando as frequências de ocorrência dos termos em documentos. Seja ***fi,j*** a frequência de ocorrência do termo ***Ki*** no documento ***dj***. A frequência total de ocorrência (*Fi*) do termo ***Ki*** na coleção é definida como o somatório das frequências de ocorrência do termo em todos os documentos:

$$F_i = \sum_{j=1}^{N} f_{i,j}$$

Onde *N* é o número de documentos na coleção, a frequência do documento do termo *Ki* é o número de documentos em que *Ki* aparece e é representado por *ni*. O valor, ou peso, de um termo ***Ki*** que ocorre em um documento ***dj*** é simplesmente proporcional à frequência termo ***fi,j***. Ou seja, quanto mais vezes um termo ***Ki*** ocorre no texto do documento ***dj***, maior é o seu peso de frequência de termo ***TFi,j***.

Essa suposição, com base na observação de que as condições de alta frequência são importantes para a descrição dos tópicos principais de um documento, leva diretamente para a seguinte formulação de peso TF:

$$tfi,j = fi,j$$

A variante do *tfi,j* usada neste trabalho é:

$$TF(t_i, d_j) = \frac{f_{i,j}}{max_i f_{i,j}}$$

Entretanto, o fator frequência do termo por si só não pode garantir o desempenho de recuperação aceitável. Especificamente, quando os termos de alta frequência não estão

concentrados em alguns documentos particulares, mas, ao invés disso, são predominantes em toda a coleção, todos os documentos tendem a ser recuperados, e isso afeta a precisão da busca. Assim, um novo fator dependente da coleta deve ser introduzido, o que favorece termos concentrados em alguns documentos de uma coleção. O fator de frequência documento inverso (*inverse document frequence* - IDF) executa essa função. O fator IDF varia inversamente com o número de documentos *n* para o qual um termo é atribuído em uma coleção de documentos *D*.

A frequência inversa do documento (IDF) foi proposta inicialmente por Jones (1972). A ideia por trás desse algoritmo é de que um termo de consulta que ocorre em muitos documentos não é um bom representante, e, logo, deve ser dado menor peso a ele do que aos que ocorrem em apenas alguns documentos. O cálculo se baseia na contagem do número de documentos na coleção em que está sendo procurado, que contêm o termo em questão.

$$idf = \frac{N}{ni}$$

Quanto mais índice se atribui a um documento, mais exaustiva a descrição se torna. A probabilidade de recuperação de um documento por uma consulta, a partir do fluxo de consulta, também aumenta. No entanto, se muitos termos são atribuídos a um documento, ele vai ser recuperado por consultas para as quais não é relevante. Isso sugere que a probabilidade de relevância de um documento recuperado é maximizada.

O número ideal de termos de índice define a exaustividade ideal para descrições de um documento. Dessa forma, os melhores termos devem ter alta frequência, mas baixa frequência nos diferentes documentos. A medida razoável da importância pode ser obtida usando o produto da frequência do termo (*tf*) e do inverso da frequência do documento (*idf*).

2.3 Modelos de Sistemas de RI

Nesta seção, serão tratados alguns aspectos técnicos sobre as principais técnicas de recuperação de informações. Para isso, será necessária a definição de alguns conceitos básicos.

Os sistemas clássicos de recuperação de informação fazem uso de uma estratégia de busca de informação baseada em uma consulta (*query*) formada pelo usuário, que serve para descrever as suas necessidades de informações, de modo a resumir as expressões-chave que caracterizam o documento que se deseja encontrar. Esses modelos consideram que cada documento é descrito por um conjunto de palavras-chave, chamadas termos de indexação.

Cada termo indexado possui associado uma palavra e um peso, que quantifica a correlação entre os termos e os documentos. Uma vez que o usuário realiza uma busca, os sistemas de RI irão comparar os termos extraídos dos documentos com a *query* fornecida, com vistas a verificar a relevância e apresentar os documentos que melhor se relacionam com a pesquisa realizada.

Para um conjunto de termos indexados em um documento, nem todos são igualmente úteis. Para Fonseca (2009), decidir sobre a importância de um termo para resumir o conteúdo de um documento não é tarefa trivial. Desconsiderando essas dificuldades, existem propriedades de um termo indexador que podem ser facilmente mensuradas e que são úteis para avaliar o potencial de um termo como tal. Por exemplo, considerando uma coleção com cem mil documentos, uma palavra que apareça em cada um desses documentos não tem utilidade alguma como termo indexador, pois ela não diz nada sobre quais documentos o usuário pode se interessar. Por outro lado, uma palavra que apareça em apenas cinco documentos torna-se bastante útil, pois ela reduz consideravelmente o espaço de documentos que podem ser de interesse do usuário.

Segundo Yates e Ribeiro Neto (1999, p. 24), "termos indexadores distintos têm relevâncias variadas quando utilizados para descrever o conteúdo de documentos. Esse efeito é capturado pela associação de pesos numéricos para cada termo indexador de um documento". Formalmente, considera-se que k_i seja um termo indexador, d_j seja um documento e $w_{i,j} >= 0$ seja um peso associado com o par (k_i, d_j). Esse peso quantifica a importância do termo indexador na descrição semântica de um documento.

Definição: considere que *t* seja o número de termos indexados no sistema e k_i seja um termo genérico. $K = \{k_1,..., k_t\}$ é o conjunto de todos os termos indexados. Um peso $w_{i,j} > 0$ é associado a cada termo k_i de um documento d_j, para um termo que não apareça no texto de um documento temos $w_{i,j} = 0$. Com o documento d_j associado a um vetor de termos indexados $\overrightarrow{d_j}$, representado por $\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, ..., w_{t,j})$. Posteriormente, consideremos g_i como sendo a função que calcula o peso associado com um termo indexado k_i em qualquer vetor t-dimensional (i.e. $g_i(\overrightarrow{d_j}) = w_{i,j}$).

Para Yates e Ribeiro Neto (1999), os pesos dos termos indexados são normalmente considerados mutuamente independentes. Isso significa que, conhecendo o peso $w_{i,j}$ associado com o par (k_i, d_j), isso não diz nada sobre o peso $w_{i+1,j}$ associado ao par (k_{i+1}, d_j). Trata-se, obviamente, de uma simplificação, porque as ocorrências de termos indexados em um documento são correlacionadas.

Considere, por exemplo, que os termos “computador” e “rede” são utilizados para indexar um dado documento que trata da área de “redes de computadores”. Frequentemente, nesse documento, o aparecimento de uma dessas duas palavras atrai o aparecimento da outra. Então, essas palavras são correlacionadas e seus pesos podem refletir essa correlação. Yates e Ribeiro Neto (1999, p. 25) ainda afirmam que “a independência mútua parece ser uma simplificação forte, mas ela também diminui a tarefa de computar os pesos dos termos indexados, possibilitando um procedimento de ranking mais rápido”.

Os modelos clássicos utilizados no processo de RI: modelo booleano, modelo vetorial e modelo probabilístico, dando maior destaque ao modelo vetorial, utilizado na implementação proposta neste trabalho.

2.3.1 Modelo Booleano

No modelo booleano, um documento é representado por um conjunto de termos indexados. As buscas são realizadas através de expressões booleanas, compostas por termos ligados através dos conectivos lógicos AND, OR e NOT (e, ou e não), e como resultado, apresentam-se os documentos que satisfaçam às restrições lógicas da expressão de busca.

2.3.2 Modelo Vetorial

O modelo Vetorial representa documentos como um vetor de termos em que cada termo possui um valor de importância associado, ou seja, um peso. A consulta feita pelo usuário, da mesma forma, é representada por um vetor. Segundo Yates e Ribeiro Neto (2011), cada consulta possui um vetor resultado, construído através do cálculo da similaridade baseada no ângulo (cosseno) entre o vetor que representa o documento e o que representa a consulta. Dessa forma, os vetores dos documentos podem ser comparados com o vetor da consulta, e o grau de similaridade entre cada um deles pode ser identificado.

Esse modelo reconhece que o “uso de pesos binários é muito limitado e propõe um *framework* pelo qual é possível mensurar uma similaridade parcial”. (YATES E RIBEIRO NETO, 1999, p. 27) Isso é alcançado pela associação de pesos não binários aos termos indexados das consultas e dos documentos. Esses pesos são utilizados, então, para medir o grau de similaridade entre cada documento armazenado no sistema e a consulta do usuário. Classificando os documentos recuperados em ordem decrescente desse grau de similaridade, o modelo vetorial leva em conta documentos que se igualam apenas parcialmente com a

consulta. O principal efeito resultante é que o conjunto de documentos ranqueados é bem mais preciso que o apresentado pelo modelo booleano.

No modelo vetorial, cada documento é representado como um vetor de termos. Cada termo possui um valor associado, que indica o grau de importância (peso - *weight*) do documento. Em outras palavras, cada documento possui um vetor associado, constituído por pares de elementos na forma {(palavra_1, peso_1), (palavra_2, peso_2),..., (palavra_n, peso_n)}.

Esses documentos podem ser organizados, por exemplo, em um arquivo invertido:

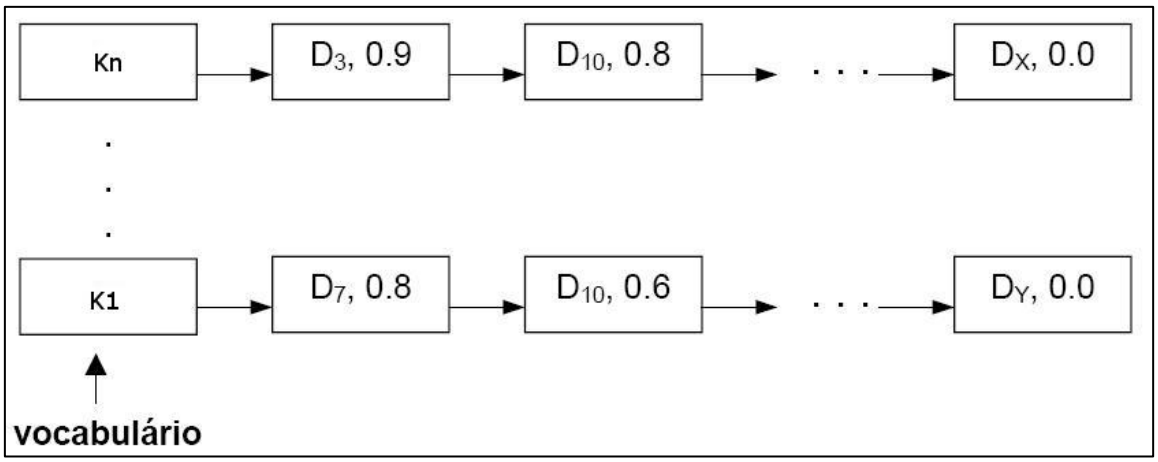


Figura 3: Documentos Indexados com os pesos de cada documento.
Fonte: Yates; Ribeiro Neto, 1999, p. 35.

Definição: para o modelo vetorial, o peso $w_{i,j}$ associado com o par (k_i, d_j) é positivo, e não binário. A seguir, também são atribuídos pesos aos termos da consulta. Considere que $w_{i,q}$ seja o associado ao par [k_i, q], onde $w_{i,q} >= 0$. Então, o vetor $\overrightarrow{q_j}$ é definido como $\overrightarrow{q_j} = (w_{1,q}, w_{2,q}, ..., w_{t,q})$, onde t é o número total de termos indexados no sistema. Assim como antes, o vetor de um documento $\overrightarrow{d_j}$ é representado por $\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, ..., w_{t,j})$.

Então, um documento d_j e uma consulta (q) do usuário q são representados como vetores t-dimensionais, como demonstrado na figura 8. Portanto, o modelo vetorial propõe avaliar o grau de similaridade de um documento d_j com a consulta q como sendo a correlação entre os vetores $\overrightarrow{d_j}$ e $\vec{q}$. Essa correlação pode ser quantificada, por exemplo, pelo cosseno do ângulo entre esses dois vetores.

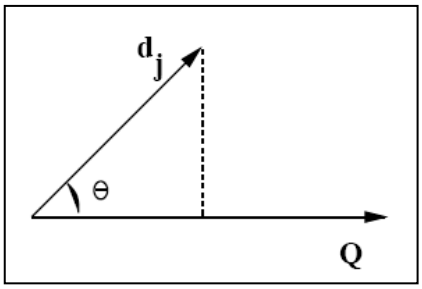


Figura 4: O cosseno do ângulo Θ é adotado como $\boldsymbol{sim(d_j, q)}$.
Fonte: Yates; Ribeiro NETO, 1999, p. 38

Então, a $sim(d_j, q)$ é calculada pela fórmula:

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|}$$

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

Partindo do princípio que $w_{i,j} \geq 0$ e $w_{i,q} \geq 0$, $sim(d_j, q)$ varia de 0 até 1. Então, ao invés de tentar dizer qual documento é relevante ou não, o modelo vetorial classifica os documentos de acordo com o seu grau de similaridade com a consulta. Um documento pode, assim, ser recuperado mesmo que ele somente atenda parcialmente a uma determinada consulta. Entretanto, para computar esse ranque, é necessário primeiro estabelecer o modo de obtenção dos pesos dos termos indexados.

Existem muitas formas de computar ranques. O trabalho de Salton e McGrill (1983) cita muitos exemplos. Como o objetivo deste trabalho não é discutir em detalhes essas técnicas, deter-se-á apenas a explicar a ideia principal por trás de uma das mais conhecidas. Os princípios a seguir dão suporte a técnicas de criação de *clusters*.

Dada uma coleção *C* de objetos e uma vaga descrição de um conjunto *A*, o objetivo de um algoritmo simples de *clustering* deve ser separar a coleção *C* em dois conjuntos: um primeiro, composto de objetos relacionados ao conjunto *A*, e um segundo, composto de objetos que não são relacionados a *A*. Uma vaga decisão quer dizer que não se tem uma informação completa para se decidir precisamente quais objetos são e quais não são do conjunto *A*. Por exemplo, alguém poderia estar buscando um conjunto *A* de carros com preços comparáveis ao de um Fusca. Sabendo que o significado comparável não é exato, não existe uma descrição precisa (única) do conjunto *A*.

Alguns algoritmos de *clustering*, conforme Salton e McGrill (1983), podem tentar separar os objetos em várias classes, de acordo com suas propriedades. Por exemplo, pacientes de um médico especializado em câncer podem ser classificados em cinco classes: terminal, avançado, em metástase, diagnosticado ou curado. Novamente, as possíveis classes continuam imprecisas e o problema é alguém decidir a qual dessas classes o paciente pertence. Portanto, será discutida apenas a versão mais simples do problema de *clustering* (i.e. a que considera a existência de apenas duas classes), por conta de que tudo que é requerido é uma decisão sobre quais documentos são relevantes e quais não são (no que diz respeito a uma dada consulta do usuário).

Salton e McGrill (1983), em um de seus trabalhos, apresentam a visão da área de RI como um problema de *clustering*. O autor considera que os documentos são uma coleção *C* de objetos e que a consulta do usuário seja uma (vaga) especificação de um conjunto *A* de objetos. Sendo assim, o problema de RI pode ser reduzido a determinar quais documentos são do conjunto A e quais não são (i.e., o problema de RI pode ser visto como um problema de *clustering*).

Segundo Yates e Ribeiro Neto (1999), em um problema de *clustering*, duas principais questões precisam ser resolvidas. Primeiro, alguém precisa definir quais características melhor descrevem os objetos do conjunto *A*. Segundo, alguém precisa determinar quais as características que melhor distinguem o conjunto *A* dos demais da coleção *C*. Do primeiro conjunto de características proveem a quantificação para a similaridade *intracluster*, enquanto o segundo provê a quantificação para a dissimilaridade *intercluster*. Os algoritmos de *clustering* que obtêm mais sucesso são aqueles que tentam balancear esses dois efeitos.

No modelo vetorial, a similaridade *intraclustering* é quantificada pela frequência crua dos termos k_i no documento d_j. Essa frequência de termos é usualmente referida como o *fator tf* (*tf-factor*) e provê uma medida de quão bem esses termos descrevem o conteúdo do documento (i.e., caracterização intradocumento). Ademais, a dissimilaridade *intercluster* é quantificada pela frequência inversa de um termo k_i entre os documentos da coleção. Esse fator é geralmente referido como a frequência inversa do documento ou o *fator idf* (*idf-factor*). A motivação para o uso do fator *idf* é que termos que aparecem em muitos documentos não são muito úteis para distinguir um documento relevante de um não relevante. Assim como um bom algoritmo de *clustering*, os esquemas mais efetivos de balanceamento de termos (atribuição de pesos) para RI tentam balancear esses dois efeitos.

Definição: considere que N seja o número total de documentos no sistema e n_i o número de documentos nos quais o termo indexado k_i aparece. Considere que $freq_{i,j}$ seja a frequência crua do termo k_i no documento d_j (i.e., o número de vezes que o termo k_i é mencionado no texto do documento d_j). Então, a frequência normalizada $f_{i,j}$ do termo k_i no documento d_j é dada por:

$$f_{i,j} = \frac{freq_{i,j}}{max_l\, freq_{j,j}}$$

Onde o máximo é computado sobre todos os termos mencionados no texto do documento d_j. Se o termo k_i não aparecer no documento d_j, então $f_{i,j} = 0$. Ademais, considere idf_i, a frequência inversa de documento para k_i, sendo dada por:

$$idf_i = \log\frac{N}{n_i}$$

Segundo Yates e Ribeiro Neto (1999), os melhores esquemas de atribuição de pesos a termos utilizam pesos atribuídos por:

$$w_{i,j} = f_{i,j} \times idf_i$$

Ou por variações dessa fórmula. Essas estratégias de atribuição de pesos são conhecidas como esquemas *tf-idf*.

As principais vantagens do modelo vetorial são:

- ✓ Seu esquema de ponderação melhora a qualidade do conjunto de respostas (das recuperações);
- ✓ Sua correspondência parcial permite a recuperação de documentos que se aproximam das condições de consulta; e
- ✓ A fórmula de ranque, baseada em cosseno, classifica os documentos de acordo com um grau de similaridade com a consulta.

Apesar de sua simplicidade, o modelo vetorial é uma estratégia de ranque bastante eficaz. Por sua popularidade, tem sido comparado com vários modelos alternativos de RI. Em geral, o modelo vetorial tem-se mostrado superior ou tão bom quanto os modelos alternativos conhecidos.

2.3.3 Modelo Probabilístico

O modelo probabilístico trata do problema da recuperação de informação de uma visão probabilística. A partir de uma consulta do usuário, o conjunto dos documentos é dividido em quatro grupos, sendo eles: o conjunto de documentos relevantes (Rel), o conjunto

dos documentos recuperados (Rec), o conjunto dos documentos relevantes recuperados (RR) e o conjunto dos documentos não relevantes e não recuperados. O que se busca nesse modelo é saber a probabilidade de um documento *D* ser ou não relevante para uma consulta. Essa informação pode ser obtida assumindo-se que a distribuição de termos na coleção seja capaz de informar a relevância provável para um documento qualquer dela.

Então, pode-se pensar no processo de busca como um procedimento de especificação das propriedades de um conjunto resposta ideal (o que é análogo à interpretação do problema de *clustering*). O problema é descobrir exatamente o que essas propriedades são. Tudo que se sabe é que elas são termos indexados dos quais a semântica pode ser utilizada para caracterizar essas propriedades.

O modelo probabilístico é baseado no seguinte pressuposto: dada uma consulta q e um documento d_j em uma coleção, o modelo probabilístico tenta estimar a probabilidade do usuário considerar o documento d_j interessante (i.e. relevante). O modelo assume que essa probabilidade de relevância depende apenas da representação da consulta do documento. Além disso, o modelo assume que existe um subconjunto de documentos que o usuário prefere definir como a resposta para a consulta q. Esse conjunto de resposta ideal é chamado de R e deve maximizar a probabilidade global de relevância para o usuário. Documentos pertencentes ao conjunto R são ditos relevantes para a consulta; documentos fora dele são ditos não relevantes. Segundo Yates e Ribeiro Neto (1999), essa pressuposição é bastante problemática porque não define explicitamente como computar as probabilidades de relevância. Na verdade, nem mesmo o espaço amostral utilizado para a definição dessas probabilidades é dado.

Dada uma consulta q, o modelo probabilístico associa a cada documento d_j, como uma medida de sua similaridade a consulta, a razão P(d_j relevante para q)/P(d_j não relevante para q) que computa as probabilidades do documento d_j ser relevante para a consulta q. Tomando as probabilidades de relevância como o ranque, minimiza-se a probabilidade de um julgamento errôneo.

Definição: para o modelo probabilístico, o peso de todos os termos indexados é binário, isto é, $w_{i,j} \in \{0, 1\}$, $w_{i,q} \in \{0, 1\}$. A consulta q é um subconjunto de termos indexados. Consideremos R como sendo o conjunto de documentos conhecidos (ou inicialmente imaginados) como relevantes. Considere $\bar{R}$ sendo o complemento de R (i.e., o conjunto de documentos não relevantes). Considere $P(R|\overrightarrow{d_j})$ com sendo a probabilidade do documento d_j

ser relevante para a consulta q e $P(\bar{R}|\overrightarrow{d_j})$ a probabilidade de d_j não ser relevante para q. então, a similaridade sim(d_j, q) do documento d_j para a consulta q é definida pela razão:

$$sim(d_j, q) = \frac{P(R|\overrightarrow{d_j})}{P(\bar{R}|\overrightarrow{d_j})}$$

A maior vantagem do modelo probabilístico, teoricamente, é que os documentos são ranqueados em ordem decrescente de suas probabilidades de serem relevantes. As desvantagens incluem: (1) a necessidade de adivinhar a separação inicial de documentos em conjuntos relevantes e não relevantes; (2) o fato de o método não considerar a frequência que um termo indexado ocorre dentro de um documento (i.e., todos os pesos são binários); e (3) a adoção da suposição de independência entre os termos indexados. Entretanto, como sugerem Yates e Ribeiro Neto (1999), não está claro se a suposição de independência entre termos indexados é uma má ideia em situações práticas.

A seguir, apresenta-se um modelo pouco citado na literatura, mas igualmente interessante: as Redes de Grafos Conceituais, que aparecem como uma alternativa com custo computacional baixo e fácil de ser implementada.

CAPÍTULO III

SISTEMAS DE RECOMENDAÇÃO

O ser humano hoje vive na era da informação, sendo os computadores e as redes de telecomunicações os grandes responsáveis por diversas mudanças socioeconômicas no século XX. Nesse cenário, deparamo-nos com uma carga de informações cada vez maior. Os slogans afirmam que o mundo está a apenas um clique de distância. Mas será que isso é verdade?

Ainda hoje há falta de informação, mas agora o problema se apresenta de forma diferente, visto que o que falta é informação relevante. A grande quantidade de conteúdo acaba gerando uma sobrecarga de informação que pode ser vista facilmente ao navegar pela Internet. Esse problema acaba causando um desestímulo no usuário, pois há dificuldade de encontrar o que se realmente deseja. Como consequência, ele se sente perdido em meio a um mundo de informações, sem saber como encontrar o que lhe interessa. Mas surge a questão: como encontrar o conteúdo de interesse do usuário em meio a tanta informação? Será que através de esforço próprio? Deve-se pedir a ajuda a alguém? Ou será que o usuário deve contar com a sorte e torcer para que encontrar o que deseja?

Os sistemas de recomendação surgiram como uma resposta à dificuldade das pessoas de realizarem escolhas em meio a grande variedade de produtos e serviços e as várias alternativas apresentadas. Segundo Ricci et al. (2011):

> O desenvolvimento dos Sistemas de Recomendação se deu a partir de uma observação bastante simples: os indivíduos muitas vezes dependem de recomendações fornecidas por outros na tomada de rotina, decisões diárias. Por exemplo, é comum que confiar na recomendação de seus amigos ao escolher um livro para ler; empregadores contam com cartas de recomendação em suas decisões de recrutamento; e ao selecionar um filme para assistir, os indivíduos tendem a ler comentários de críticos de cinema para auxiliar em suas escolhas. (RICCI et al., 2011, p. 2)

Sistemas de Recomendação são ferramentas de software e técnicas que fornecem sugestões de itens que sejam úteis para um usuário. Segundo Ricci et al. (2011), esses sistemas tentam prever quais são os produtos ou serviços mais adequados, com base nas preferências e restrições do usuário. Para completar essa tarefa, eles coletam dos usuários suas preferências, que podem ser explicitas, através das avaliações de produtos, ou implícitas, inferidas através da interpretação das ações do usuário.

Ricci et al. (2011) apontam os sistemas de *e-commerce* como sendo um dos grandes responsáveis pela evolução dos Sistemas de Recomendação.

> Como o desenvolvimento dos sistemas de *e-commerce*, uma necessidade emergente surgiu para fornecer recomendações derivadas de filtrar toda a gama de alternativas disponíveis. Usuários foram encontrando muita dificuldade para chegar às escolhas mais adequadas a partir da imensa variedade de itens (produtos e serviços) que esses sites estavam oferecendo. (RICCI et al., 2011, p. 3)

Os Sistemas de Recomendação surgiram para auxiliar no processo de indicar e receber indicações. Dessa forma, procuram facilitar a busca por conteúdo interessante ao usuário, prevendo itens que possam ser relevantes ao mesmo. Por assim ser, nos últimos anos, os Sistemas de Recomendação têm se mostrado um meio valioso para lidar com o problema da sobrecarga de informações. Esses sistemas apontam o usuário no sentido de novas informações, ainda não exploradas, que possam ser de interesse para as suas tarefas.

Esses sistemas geram recomendações usando vários tipos de conhecimentos e dados sobre o usuário, tais como itens disponíveis e transações anteriores armazenadas em bancos de dados personalizados. Essas recomendações podem ser úteis ou não para o usuário. Para descobrir a relevância de uma sugestão, o usuário pode fornecer, imediatamente ou numa etapa seguinte, um retorno implícito ou explícito. Todas essas ações do usuário e feedbacks podem ser armazenados no banco de dados (perfil) e utilizados para a geração de novas recomendações nas interações próximas do sistema do usuário.

Segundo Ricci et al. (2011), a fim de implementar a sua função principal, identificando os itens úteis para o usuário, o sistema de recomendação deve predizer que item vale a pena recomendar. Para tanto, o sistema deve ser capaz de prever a utilidade de alguns deles, ou, pelo menos, comparar a utilidade de alguns dos itens, e, então, decidir quais recomendar, com base nessa comparação.

> Na sua forma mais simples, recomendações personalizadas são oferecidas como listas de classificados de itens. Na realização deste ranking, os Sistemas de Recomendação tentam prever quais os produtos ou serviços mais adequados são, com base nas preferências e restrições do usuário. A fim de completar uma tarefa tão computacional, os Sistemas de Recomendação coletam dos usuários suas preferências, e que são explicitados, por exemplo, como notas de produtos, ou são inferidas por interpretar as ações do usuário. Por exemplo, um Sistema de Recomendação pode considerar a navegação para uma página de determinado produto como um sinal implícito de preferência para os itens mostrados na página. (RICCI et al., 2011, p. 2)

Como já mencionado, o estudo de Sistemas de Recomendação é relativamente novo em comparação com a investigação sobre outras ferramentas clássicas de sistemas de

informação e técnicas (por exemplo, bases de dados ou motores de busca). Sistemas de Recomendação surgiram como uma área de pesquisa independente, em meados da década de 1990. Segundo Ricci et al. (2011), o interesse nesses sistemas aumentou dramaticamente nos últimos anos, como os seguintes fatos indicam:

1. Sistemas de Recomendação desempenham um papel importante em sites da Internet altamente cotados, como Amazon.com, YouTube, Netflix, Yahoo, Tripadvisor, Last.fm, e IMDb. Além disso, muitas empresas de mídia estão desenvolvendo e implantando Sistemas de Recomendação como parte dos serviços que prestam aos seus assinantes. E o caso do Netflix, o serviço de aluguel de filmes online, premiado com um prêmio de um milhão de dólares para a equipe que primeiro conseguiu melhorar substancialmente o desempenho do seu Sistema de Recomendação.
2. Há conferências e workshops relacionados com a área, como, especificamente, o ACM Sistemas de Recomendação (RecSys), criado em 2007 e, agora, o evento anual premier em pesquisa e tecnologia de aplicações recomendadoras. Além disso, as sessões dedicadas aos Sistemas de Recomendação são frequentemente incluídas nas conferências mais tradicionais na área de bases de dados, sistemas de informação e sistemas adaptativos. Entre essas conferências são dignos de menção ACM SIGIR Grupo de Interesse Especial em *Information Retrieval* (SIGIR), modelagem de usuários, adaptação e personalização (UMAP), e Grupo de ACM Interesse Especial sobre Gestão de Dados (SIGMOD).
3. Nas instituições de ensino superior em todo o mundo, cursos de graduação e pós-graduação estão agora inteiramente dedicados aos Sistemas de Recomendação. Tutoriais sobre esses sistemas são muito populares no computador, bem como conferências científicas.
4. Houve várias edições especiais em revistas acadêmicas que cobrem pesquisas e desenvolvimentos no campo Sistemas de Recomendação. Entre as revistas que têm dedicado a esses sistemas estão: AI Communications (2008); Sistemas Inteligentes IEEE (2007); Jornal Internacional de Comércio Eletrônico (2006); International Journal of Science and Applications Computer (2006); ACM Transactions em Interação Humano-Computador (2005); e ACM Transactions em Sistemas de Informação (2004).

Ricci et al. (2011) destacam algumas razões pelas quais os prestadores de serviços podem querer explorar essa tecnologia, entre elas:

1. **Aumentar o número de itens vendidos.** Essa é, provavelmente, a função mais importante para um Sistema de Recomendação comercial, ou seja, um sistema capaz de vender um conjunto adicional de itens em comparação com aqueles normalmente vendidos sem qualquer tipo de recomendação. Esse objetivo é alcançado porque os itens recomendados são susceptíveis de satisfazer as necessidades do usuário. Aplicações não comerciais têm objetivos semelhantes, mesmo não havendo nenhum custo para o usuário, que está associado com a seleção de um item. Por exemplo, uma rede de conteúdo visa a aumentar o número de itens de notícias lidas em seu site. Em geral, pode-se dizer que, do ponto de vista do fornecedor de serviços, o objetivo principal de introdução de um Sistema de Recomendação é aumentar a taxa de conversão, ou seja, o número de usuários que aceitam a recomendação e consomem um item, em comparação com o número de simples visitantes que apenas percorrem as informações.
2. **Vender mais itens diversos**. Outra importante função de um Sistema de Recomendação é permitir que o usuário selecione itens que podem ser difíceis de encontrar, sem uma recomendação precisa. Por exemplo, em relação a filmes, os Sistemas de Recomendação, tais como o Netflix, o prestador de serviços está interessado em alugar todos os DVDs no catálogo, e não apenas os mais populares. Isso pode ser difícil sem um Sistema de Recomendação, desde que o prestador de serviços não pode arcar com o risco dos filmes não servirem ao gosto de um determinado usuário. Portanto, o Sistema de Recomendação sugere ou anuncia filmes impopulares para os usuários certos.
3. **Aumentar a satisfação do usuário**. Um Sistema de Recomendação bem concebido também pode melhorar a experiência do usuário com o site ou aplicativo. O usuário encontrará as recomendações relevantes e, com uma interação humano-computador bem projetada, poderá vir a gostar de usar o sistema. A combinação de recomendações eficazes, exatas, e uma interface utilizável irá aumentar a avaliação subjetiva do utilizador do sistema. Este, por sua vez, irá aumentar o uso do sistema e a probabilidade de que as recomendações sejam aceitas.

4. **Aumentar a fidelidade do usuário**. Um usuário deve ser leal a um site que, quando visitou, reconheceu o antigo cliente e, assim, tratá-lo como um visitante valioso. Essa é uma característica regular de um Sistema de Recomendação, já que muitos desses sistemas realizam as recomendações baseados em informações adquiridas a partir do usuário em interações anteriores, por exemplo, suas classificações de itens. Consequentemente, quanto mais tempo o usuário interage com o site, mais refinado seu modelo de usuário se torna, ou seja, a representação do sistema de seus preferências.
5. **Entender melhor o que o usuário deseja**. Outra função importante dos Sistemas de Recomendação, que podem ser aproveitados para muitas outras aplicações, é a descrição de preferências do usuário, recolhidas de forma explícita ou prevista pelo sistema. O prestador de serviços pode, em seguida, decidir voltar a usar esse conhecimento para uma série de outras metas, como a melhoria da gestão de estoque ou produção do item. Por exemplo, no domínio do turismo, as organizações de gestão de destinos podem decidir anunciar uma região específica a novos setores de clientes ou anunciar um determinado tipo de mensagem promocional derivada através da análise dos dados coletados pelos Sistemas de Recomendação.

Sistemas de Recomendação envolvem a construção de um modelo ou perfil de interesses do usuário, o que varia de acordo com a técnica de recomendação utilizada, sendo as principais delas a Filtragem Colaborativa, a Recomendação Baseada em Conteúdo, o Modelo Híbrido, entre outras. A seguir, são detalhados esses três modelos.

3.1 Filtragem Baseada em Conteúdo

O problema de recomendação de itens foi estudado extensivamente, e dois paradigmas principais emergiram. Os Sistemas de Recomendação Baseada em Conteúdo, que tentam recomendar itens semelhantes aos que um determinado usuário tenha gostado no passado; e os Sistemas de Recomendação Colaborativa, que identificam os usuários cujas preferências são semelhantes aos do usuário dado e recomendam itens que tenham gostado. (BALABANOVIC, 1997 apud RICCI et al., 2011)

Como já apresentado no tópico acima, os Sistemas de Recomendação têm o propósito de orientar os usuários, de forma personalizada, para objetos interessantes, em um grande espaço de opções possíveis. Os Sistemas de Recomendação baseados em conteúdo

tentam recomendar itens semelhantes a aqueles que um determinado usuário tenha gostado no passado. Para Ricci et al. (2011),

> [...] o processo básico realizado por um recomendador baseada em conteúdo consiste em combinar os atributos de um perfil de usuário em que as preferências e interesses são armazenados, com os atributos de um objeto de conteúdo (item), a fim de recomendar ao usuário novo interessante itens. (RICCI et al., 2011, p. 73)

Burke (2002 apud RICCI et al., 2011, p. 74) afirma que "os Sistemas de Recomendação têm o efeito de guiar os usuários de forma personalizada para objetos interessantes ou úteis em um grande espaço de opções possíveis".

> A abordagem dos Sistemas de Recomendação Baseado em Conteúdo analisa um conjunto de documentos e/ou descrições de itens previamente avaliados por um usuário, e construí um modelo ou perfil dos interesses dos utilizadores com base nas características dos objetos avaliados por esse usuário. O perfil é uma representação estruturada dos interesses dos utilizadores adotada para indicar novos itens interessantes. O processo de recomendação consiste basicamente em combinar os atributos do perfil do usuário contra os atributos de um objeto de conteúdo. O resultado é um julgamento de relevância que representa o nível do usuário de interesse por esse objeto. Se um perfil reflete com precisão as preferências do usuário, é de tremenda vantagem para a eficácia de um processo de acesso à informação. Por exemplo, poderia ser usado para filtrar os resultados de pesquisa por decidir se um usuário está interessado em uma página da Web específica ou não e, em caso negativo, impedindo que ele seja exibido. (RICCI et al., 2011, p. 75)

Tendo as suas origens na área de recuperação de informação, a filtragem baseada em conteúdo parte do princípio de que os usuários tendem a se interessar por itens similares aos que tenham interessado no passado. Essa abordagem tem como objetivo gerar, de forma automática, descrições dos conteúdos dos itens e comparar com os interesses do usuário para verificar a sua relevância.

Nesse modelo, o conteúdo é geralmente descrito por palavras-chave, e a ponderação desses termos, ou seja, o cálculo da importância dessas palavras pode ser definido de diferentes formas, sendo que, geralmente, dá-se através do modelo estatístico, proposto por Salton e Buckley (1988), TF-IDF. A aplicação dessas técnicas de indexação permite mensurar a relevância de um termo para o documento. Feito isso, é estabelecida a comparação do perfil do usuário para verificar a similaridade, ou seja, se é relevante a ele. O perfil pode ser obtido de duas maneiras: explicitamente, quando o usuário responde a questionários e, a partir de suas respostas, o sistema pode obter os seus interesses; ou implicitamente, quando o sistema, de maneira automática, verifica suas preferências, de acordo com os itens visualizados.

Segundo Ricci et al. (2011), a maioria dos Sistemas de Recomendação Baseados em Conteúdo usam modelos de recuperação relativamente simples, como a comparação de palavras-chave, ou o modelo de espaço vetorial, com base na ponderação de termos TF*IDF.

3.1.1 Arquitetura

Para Ricci et al. (2011), os Sistemas Filtragem de Informação Baseada em Conteúdo precisam de técnicas adequadas para representar os itens e produzir o perfil do usuário, bem como algumas estratégias para comparar o perfil do usuário com a representação item. A arquitetura de alto nível de um Sistema de Recomendação Baseada em Conteúdo está representado na figura 5. O processo recomendação é realizado em três passos, sendo cada um deles controlado por um componente separado:

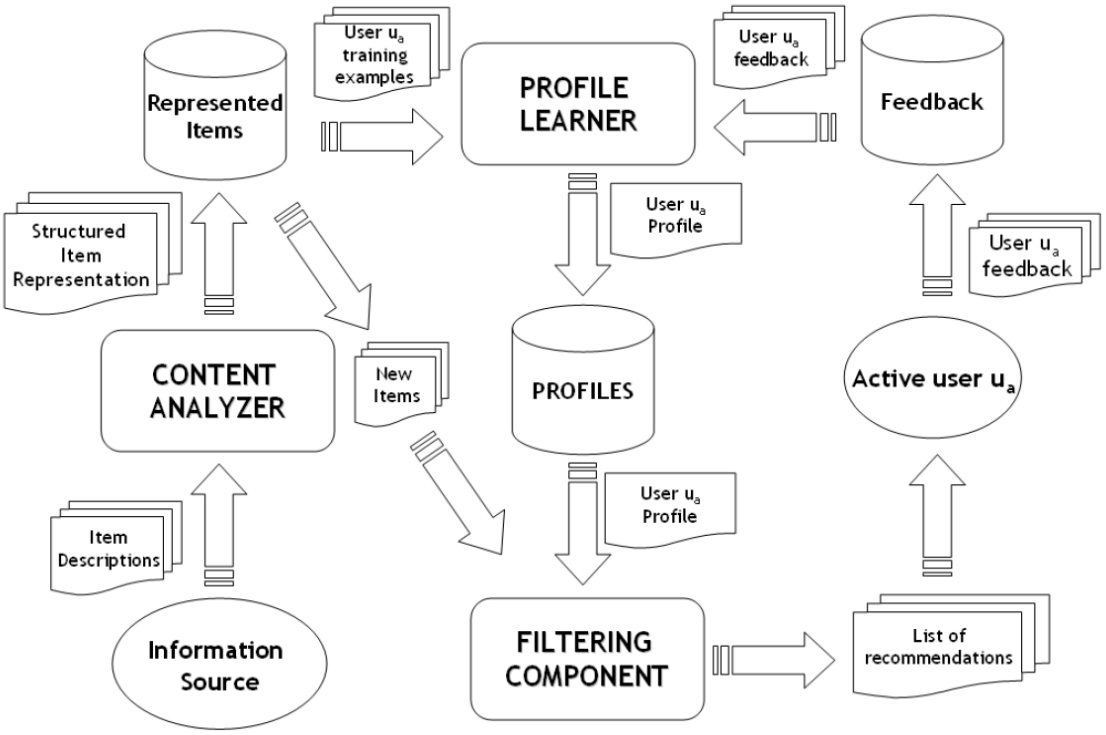


Figura 5: Arquitetura de Alto nível de um Sistema de Recomendação Baseado em Conteúdo.
Fonte: Ricci et al., 2011, p. 76.

- **Content Analyzer** (Analisador de Conteúdo): Quando a informação não tem estrutura (por exemplo, texto), uma espécie de etapa de pré-processamento é necessária para extrair informação estruturada relevante. A principal responsabilidade do componente é representar o conteúdo de itens (por exemplo, documentos, páginas da Web, notícias, descrições de produtos, e

outros) provenientes de fontes de informação em uma forma apropriada para as próximas etapas de processamento. Os itens de dados são analisados por meio de técnicas de extração de características, a fim de mudar a representação item do espaço de informação original para o destino (por exemplo: páginas Web representados como vetores de palavras-chave). Essa representação é a entrada para o PROFILE LEARNER e FILTERING COMPONENT.

- **Profile Learner:** Esse módulo recolhe dados representativos das preferências do usuário e tenta generalizá-los, a fim de construir o perfil do mesmo. Normalmente, a estratégia de generalização é realizada por meio de técnicas de aprendizagem de máquina, que são capazes de inferir um modelo de interesses do usuário a partir de itens que ele tenha gostado ou não no passado.
- **Filter Component:** Esse módulo explora o perfil do usuário para sugerir itens relevantes, combinando a representação contra esse perfil de itens a serem recomendados. O resultado é um julgamento de relevância binária ou contínua (calculado usando algumas métricas de similaridade), o último caso, resultando em uma lista ordenada de itens potencialmente interessantes.

3.1.2 Vantagens em Relação à Filtragem Colaborativa

Para Ricci et al. (2011), a adopção do paradigma de Recomendação Baseada em Conteúdo tem várias vantagens quando comparada com a Filtragem Colaborativa, como, por exemplo:

- **Independência do Usuário** - recomendação baseada em conteúdo para explorar exclusivamente ratings fornecidos pelo usuário ativo para construir o seu próprio perfil. Ao invés disso, os métodos de filtragem colaborativa precisam de avaliações de outros usuários, a fim de encontrar os "vizinhos mais próximos" do usuário ativo, ou seja, os usuários que têm gostos semelhantes, uma vez que avaliaram os mesmos itens da mesma forma. Em seguida, apenas os itens que são mais apreciados pelos vizinhos do usuário ativo serão recomendados.
- **Transparência** - Explicações sobre como funciona o sistema de recomendação pode ser fornecido listando explicitamente recursos de conteúdo ou descrições que causaram um item para ocorrer na lista de recomendações. Essas

características são indicadores para consultar, de modo a decidir se confia em uma recomendação. Por outro lado, sistemas colaborativos são caixas pretas, já que a única explicação para um item de recomendação é que os usuários desconhecidos, com gostos similares, gostaram dele.

- **Novo Item** – A recomendação baseada em conteúdo é capaz de recomendar itens ainda não avaliados por qualquer usuário. Como consequência, eles não sofrem com o problema primeiro-avaliador, o que afeta a filtragem colaborativa, que conta somente com as preferências dos usuários para fazer recomendações. Portanto, até que o novo item seja avaliado por um número significativo de usuários, os sistemas não são capazes de recomendá-lo.

3.2 Filtragem Colaborativa

Na tentativa de imitar o comportamento de um usuário, pedindo recomendação a outras pessoas, o primeiro Sistema de Recomendação aplicou algoritmos para alavancar recomendações produzidas por uma comunidade de usuários, com vistas a fornecer recomendações para um usuário ativo, ou seja, um usuário procurando sugestões. As recomendações foram para os itens que os usuários semelhantes (aqueles com gostos parecidos) gostaram. Essa abordagem é chamada de Filtragem Colaborativa e sua lógica é que, se o usuário ativo concordou, no passado, com alguns usuários, em seguida, outras recomendações provenientes desses usuários similares devem ser também muito importantes e de interesse para o usuário ativo.

Segundo Jannach et al. (2011, p. 3), “a ideia básica desses sistemas é que, se os usuários compartilhavam os mesmos interesses no passado - se eles viram ou compraram os mesmos livros, por exemplo - eles também terão gostos semelhantes no futuro”. Assim, se, por exemplo, o usuário A e o usuário B têm um histórico de compras que se sobrepõem fortemente, e o usuário A, recentemente, comprou um livro que B ainda não tenha visto, a lógica básica é a de propor esse livro também para B. Essa seleção de possíveis livros interessantes envolve filtrar os mais promissores a partir de um conjunto grande, de modo que os usuários possam, implicitamente, colaborar uns com os outros. Essa é técnica chamada de Filtragem Colaborativa.

Diferentemente da filtragem baseada em conteúdo, a Filtragem Colaborativa tenta predizer a avaliação de um item para o usuário com base nas avaliações de outros usuários com o perfil semelhante a ele. Nessa abordagem, o usuário preenche um questionário em que

avalia diversos itens. A partir de suas avaliações, são criados grupos de usuários que apresentam perfis semelhantes, que deram notas semelhantes aos mesmos itens. Sendo assim, uma vez que esse grupo de usuário tenha avaliado com uma boa nota certo item que um usuário ainda não tenha avaliado, é pressuposto que o usuário achará este item relevante, haja vista que, no passado, seus gostos se assemelharam bastante. Alag e Macmanus (2009) denominam inteligência coletiva a ideia de utilizar "gostos" de um grupo de pessoas para fazer recomendações para outras.

3.3 Híbrido

A abordagem de filtragem híbrida busca superar as limitações individuais da filtragem colaborativa e da filtragem baseada em conteúdo. Para isso, ela combina os pontos fortes de ambas para criar um sistema que possa atender melhor as necessidades do usuário.

Já vimos que as diferentes abordagens discutidas até agora têm certas vantagens e, é claro, desvantagens, dependendo da configuração problema. Uma solução óbvia é a de combinar técnicas diferentes para gerar recomendações melhores ou mais precisas. Se, por exemplo, existe o conhecimento da comunidade e informações detalhadas sobre os itens individuais, um sistema de recomendação pode ser reforçado por hibridação de filtragem colaborativa com técnicas baseadas em conteúdo.

3.4 Recuperação de Informação VS Recomendação de Informação

Ambos os Sistemas de Recuperação e Recomendação de Informação lidam com um grande volume de informação. Mas qual a diferenças dessas duas tecnologias? Como já apresentado, os Sistemas de Recuperação de Informação surgiram a partir da necessidade de se obter informações em um curto espaço de tempo. Sendo assim, criaram técnicas de organização das informações, estratégias de busca e mecanismos de busca. A Recuperação de Informação trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, bem como qualquer sistema, técnicas ou máquinas que são empregadas para realizar essas operações. (MOOERS, 1951)

Os Sistemas de Recomendação de Informação foram uma consequência da grande quantidade de conteúdo, o que acabava gerando uma sobrecarga de informação. Surgiram como uma resposta à dificuldade das pessoas realizarem escolhas, em meio a uma grande variedade de produtos, serviços e as várias alternativas apresentadas. Por assim ser, os

Sistemas de Recomendação são ferramentas de software e técnicas que fornecem sugestões de itens que sejam úteis para um usuário. Segundo Ricci et al. (2011), esses sistemas tentam prever quais são os produtos ou serviços mais adequados, com base nas preferências e restrições do usuário.

Recuperação da Informação

- Usuário descreve a sua necessidade de informação, através de uma consulta (query);
- Casamento da consulta com os documentos armazenados;
- Interação provocada pelo usuário;
- Baseia-se na percepção de uma necessidade do momento.

Recomendação de Informação

- Abordagem distinta;
- Mantém um perfil dos interesses dos usuários;
- Refere-se às preferências dos usuários;
- Entrega de informações para as pessoas que realmente necessitam.

O fato de este trabalho fazer uso de um motor de busca não significa que o sistema proposto seja apenas um novo sistema de Recuperação de Informação. Mais que isso, o sistema obtém informações do perfil do usuário de maneira automática e implícita, sem a ação direta do mesmo nesse processo. Além disso, uma vez que os links são recomendados, o usuário irá avaliá-los, e essa avaliação irá compor o seu perfil para futuras recomendações personalizadas.

CAPÍTULO IV

MINERAÇÃO DE TEXTO

Os avanços tecnológicos na área de compartilhamento e armazenamento de dados fizeram com que o volume de informações no formato digital crescesse em proporções antes inimagináveis. Segundo Kuechler (2007), 80% desses dados não estão em formato estruturado, sendo que uma grande parte deles são textos. Essas informações incluem: e-mails, arquivos eletrônicos gerados por software, editores de texto, páginas web, campos textuais em banco de dados, e mais. Entretanto, esse formato foi criado para que os documentos pudessem ser visualizados por seres humanos, não sendo adequado para a manipulação das informações nele contidas por sistemas computacionais.

Em geral, esses conteúdos são muito relevantes para as organizações, pois, segundo Han e Kamber (2006), constituem um importante repositório organizacional, o que envolve o registro de histórico de atividades, memorandos, documentos internos, e-mails, projetos, estratégias e o próprio conhecimento adquirido. Wives (2002) afirma que esse tipo de informação é muito importante para que os empresários consigam identificar novos dados e conhecimentos que estejam, de alguma forma, implícitos ou escondidos nos seus Sistemas de Informação, e que não possam ser recuperados pelos meios tradicionais oferecidos por eles.

Para Rezende, Marcacini e Moura (2011), a organização inteligente dessas coleções textuais é de grande interesse para a maioria das instituições, pois agiliza os processos de busca e recuperação da informação. Entretanto, o volume de dados textuais armazenados extrapola a capacidade humana de, manualmente, analisá-los e compreendê-los por completo. Sendo assim, a mineração de textual surgiu em decorrência da necessidade de se descobrir, de forma automática, informações em documentos nos quais, segundo Aranha e Passos (2006), o uso dessa tecnologia permite recuperar informações, extrair dados, resumir, descobrir padrões, associações, regras e realizar análises qualitativas ou quantitativas em documentos de texto. A mineração de textos, também chamada de mineração de dados textuais, permite transformar grande parte desses conteúdos não estruturados em conhecimento útil para as organizações.

Konchady (2006) apresenta uma definição geral de mineração de texto, que inclui todos os tipos de processamento de texto que tratam de encontrar, organizar e analisar

informações. Lopes (2004), por sua vez, apresenta o seguinte conceito: Text mining, também conhecido como *Text data mining* ou *Knowledge Discovery from textual databases*, refere-se ao processo de extrair padrões interessantes e não triviais ou conhecimento a partir de documentos em textos não estruturados. *Text mining* pode também ser definido como um conjunto de técnicas e processos que se prestam a descobrir conhecimento inovador nos textos. Essa nova tecnologia está sendo empregada, atualmente, em projetos de diversas áreas.

Morais e Ambrósio (2007) apresentam a mineração de texto como um processo de descoberta de conhecimento que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, as quais, normalmente, não poderiam ser recuperadas por meio de métodos tradicionais de consulta, pois as informações contidas nesses textos não pode ser obtidas de forma direta, uma vez que, em geral, estão armazenadas em formato não estruturado.

Dessa forma, podemos entender a mineração de texto como sendo a aplicação de um conjunto de diferentes técnicas em dados não estruturados com o objetivo de obter informações, as quais, muitas das vezes, podem não estar presentes de maneira explícita nesses documentos. Inspirado pelo *Data Mining* ou mineração de dados, que procura descobrir padrões emergentes em banco de dados estruturados, a mineração de textos tem como objetivo a extração de conhecimento úteis em dados não estruturados ou semiestruturados.

Para Konchady (2006), tanto mineração de dados quanto mineração de texto buscam por informações escondidas e empregam algoritmos semelhantes de Inteligência Artificial, aprendizagem de máquina, e estatística. Mas, enquanto a mineração de dados lida com dados estruturados, mineração de texto lida com dados não estruturados. Isto é, uma extensão da área de Data Mining, focada na análise de textos.

Wives (2002) apresenta o processo de descoberta de conhecimento em textos como uma evolução natural da recuperação de informações, já que os sistemas de recuperação de informação passaram a adotar algumas técnicas de análise de informações e de aprendizado de máquina, muitas das quais provenientes da área de descoberta de conhecimento em bases de dados. Assim, ao invés do usuário ter que analisar quais dos documentos retornados são realmente relevantes, o próprio sistema faz essa análise e retorna as informações de forma condensada e resumida.

As técnicas empregadas na mineração de textos incluem: extração de informação, sumarização, agrupamento de textos (*clustering*), categorização (ou classificação), entre outras.

4.1 Extração de Informação

Segundo Scarinci (1997), a extração de informação surgiu dentro da área de Processamento de Linguagem Natural (PLN), possuindo muitos componentes desse sistema, sendo que muito deles também são utilizados em sistemas de recuperação de informação. Por esse motivo, o processo de extração se torna muito parecido com o processo de indexação de informações. A extração de informação identifica, dentro de um documento textual, trechos que correspondem a dados relevantes para o usuário. Em geral, ela envolve a identificação de padrões que representam um contexto-chave dentro do texto.

Cowie (1996) faz uma interessante comparação para explicar a relação e os objetivos dos sistemas de recomendação e extração de informações. Para Cowie (1996), os sistemas de recuperação de informação podem ser vistos como "colheitadeiras" que devolvem material útil de um vasto campo de materiais brutos. Com grandes quantidades de informações potencialmente úteis em mãos, um sistema desse tipo pode, então, transformar o material bruto, refinando-o e reduzindo-o à ideia do texto original.

Se por um lado a área de mineração textual faz fronteira com a mineração de dados**,** por outro lado também interage com a área de recuperação de informações. Konchady (2006) apresenta bem essa diferença quando afirma que, enquanto os sistemas de recuperação de informação lidam com o problema de encontrar documentos relevantes em uma coleção de documentos, a extração de informação busca identificar textos relevantes em um documento. Uma informação útil é definida como um segmento de texto e seus atributos associados.

4.2 Sumarização

Sumarização é uma técnica que busca identificar as palavras e frases mais importantes de um documento e gerar um resumo ou sumário. Esse sumário serve para dar uma visão geral do documento, salientando as partes mais importantes. Dessa forma, o usuário pode identificar rapidamente o assunto abordado em um documento, sem a necessidade de ter que ler todo o conteúdo na íntegra.

Uma das tarefas de mineração de texto é tornar mais fácil o processo de encontrar informações relevantes em uma coleção de documentos. Segundo Konchady (2006), o objetivo do resumo é transmitir com precisão a essência de um documento com o mínimo de texto possível, ou permitir que um leitor decida se um documento deve ser lido na íntegra. Dessa forma, a sumarização reduz o tempo que um pesquisador deve passar navegando para localizar um texto útil.

4.3 Agrupamento (*clustering*)

O agrupamento é uma estratégia utilizada para agrupar documentos semelhantes. Diferentemente da categorização, em que as classes devem ser previamente modeladas ou descritas, o *Clustering* agrupa os documentos em tempo real. Para que isso ocorra, esta técnica identifica os documentos com assunto similar e os aloca em um grupo. Assim, gera agrupamentos de documentos similares. Nessa técnica, não há a necessidade de se ter conhecimento prévio sobre os assuntos dos documentos ou do contexto dos mesmos, pois eles são descobertos automaticamente pelo processo de agrupamento.

4.4 Categorização e Classificação

A classificação é uma técnica que utiliza o conteúdo do documento para identificar a que classe ou categoria ele pertence. Para tanto, é necessário que as classes sejam previamente modeladas ou descritas através de suas características, atributos ou fórmula matemática. Segundo Konchady (2006), a tarefa de classificação dos documentos em categorias pré-definidas é importante para construir um diretório de informação útil. Muitas vezes, uma coleção fornece documentos em um diretório de temas, além de uma interface de busca.

CAPÍTULO V

DESCRIÇÃO DO SISTEMA

Como se pode perceber, todas as tecnologias apresentadas neste trabalho estão voltadas para a manipulação, organização e busca de informação. Mas não de qualquer tipo de informação, e sim daquelas relevantes e úteis para o usuário.

O surgimento e a expansão da Internet na década de 1990 foi responsável por profundas mudanças na EAD, principalmente no que diz respeito ao uso de TICs, que trouxeram um novo formato a essa modalidade de ensino, por meio do qual alunos e professores podem interagir entre si. Essas mudanças permitiram uma nova dinâmica, o que, segundo Piva (2011), provocou uma mudança de paradigma no sentido de que a individualização cedeu lugar à colaboração e a aprendizagem independente passou a ser sustentada por experiências colaborativas entre alunos e professores e alunos entre si.

Uma das principais ferramentas dos AVA voltadas para a construção do conhecimento de forma colaborativa são os fóruns de discussão, que são espaços de discussões e troca de ideias em torno de temas propostos por seus participantes. Esse instrumento permite que cada participante submeta sua colaboração referente ao tema proposto, buscando, assim, entendimento mútuo. Segundo Silva (2006 apud OKADA, 2006), o fórum é uma ferramenta assíncrona que representa um espaço para debates, no qual pode ocorrer o entrelaçamento de muitas vozes para a construção e desconstrução de pensamentos, para questionar e responder dúvidas, trilhando novos caminhos para a aprendizagem.

Sobre a importância desses debates, Kenski (2002) tece o seguinte comentário:

> Interagir com o conhecimento e com as pessoas para aprender é fundamental. Para a transformação de um determinado grupo de informações em conhecimentos é preciso que estes sejam trabalhados, discutidos, comunicados. As trocas entre colegas, os múltiplos posicionamentos diante das informações disponíveis, os debates e as análises críticas auxiliam a sua compreensão e elaboração cognitiva. As múltiplas interações e trocas comunicativas entre parceiros do ato de aprender possibilitam que estes conhecimentos sejam permanentemente reconstruídos e reelaborados. (KENSKI, 2002, p. 258)

Com relação à participação e envolvimento nos fóruns de discursão, Oliveira (2005) salienta que a participação no espaço criado pelo fórum pede preparo, o que geralmente é provido por meio de leituras adequadas, pesquisas, resgates ao *background*

próprio a cada participante, entre outras formas de busca. Trata-se de organizar o pensamento, enriquecendo-o com pertinentes referências, permitindo o uso do espaço de discussões e reflexões proporcionado pelo fórum para gerar colaborações, para agregar ideias. É a chance de valorizar o conhecimento abalizado, com espaço para opiniões pessoais – e a discussão das mesmas – sem que essa iniciativa represente uma apologia ao "achismo" e ao acúmulo de debates improfícuos, destituídos de solidez teórica. O tempo comunicacional, assíncrono, favorece semelhante postura em um espaço potencialmente livre de conflitos.

Partindo desse problema, este trabalho propõe a utilização de um sistema com o objetivo de fomentar discussões em fóruns através da recomendação de links.

5.1 Arquitetura

Para que seja possível realizar a recomendação de links, deve-se fazer uso de uma ferramenta de mineração de texto para a prévia identificação das palavras- chave que representam os tópicos do fórum. Uma vez identificas, essas palavras são submetidas a um motor de busca, que retornará os links que poderão oferecer novos conteúdos ao usuário para estar se aprofundando ainda mais no contexto da discussão. Esses links serão avaliados pelo usuário, e alimentarão um perfil que, no futuro, fará novas sugestões baseadas em seus gostos.

A imagem abaixo apresenta o fluxograma representativo da arquitetura do sistema proposto.

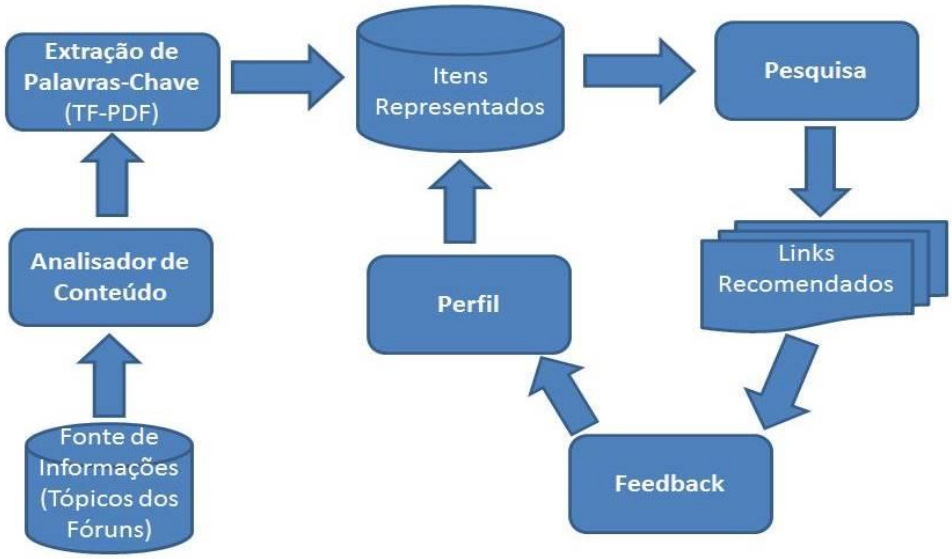


Figura 6: Arquitetura do Sistema.
Fonte: O autor.

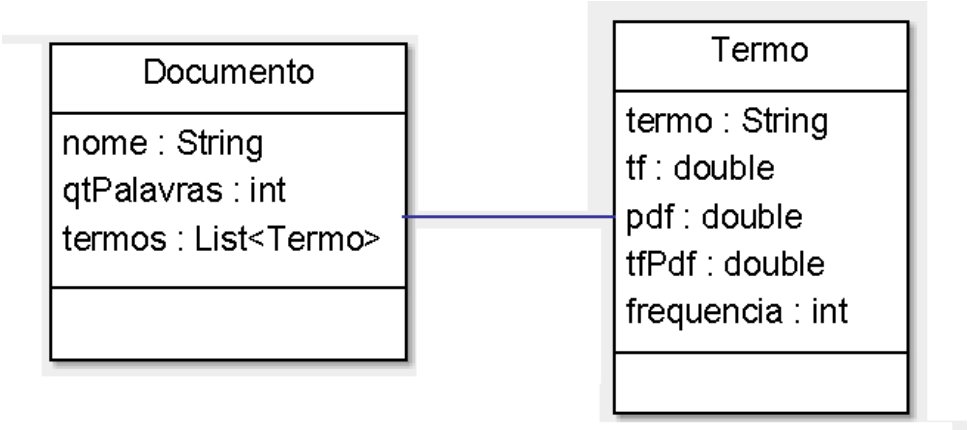
O sistema proposto neste trabalho busca recomendar links com o objetivo de promover discussões em fóruns de um AVA, com o propósito de oferecer novas possibilidades de leituras, de modo a propiciar o desenvolvimento novas ideias e o aprimoramento do conhecimento.

Segundo Oliveira (2005), a participação nos fóruns exige um preparo, geralmente provido através de pesquisas e leituras adequadas, em que cada participante enriquece o debate com referências pertinentes, não havendo espaço para achismo. Dessa forma, o uso desse sistema tem como objetivo abrir novas possibilidades de leitura de tópicos relacionados aos temas debatidos, estimulando os alunos a irem além do material disponibilizado no AVA.

A dinâmica de funcionamento do sistema implementado durante esta pesquisa está disposta da seguinte maneira:

1. O processo de recomendação de links tem início através da análise e do processamento das informações. Nesse caso, a fonte de informações são os tópicos (discussões) do fórum.

Cada tópico do fórum é representado em um arquivo de texto (.txt), cujo conteúdo será representado no sistema através da classe documento, composta pelos atributos: nome, quantidade de palavras e uma lista de termos. Para representar cada termo contido no documento, foi criada a classe termo. Esta, por sua vez, contém os atributos: termo; a frequência do termo (*tf*, proposto por Salton e Buckley (1988) e explicado em mais detalhe logo a frente); a frequência ponderado do termo (*pdf*, proposto por Ishizuka (2001) e explicado em mais detalhes mais adiante); o peso do termo, calculado através da multiplicação da frequência do termo pela frequência ponderado do termo(*tf-pdf*); e o número de ocorrências do termo.



2. O analisador de conteúdo consiste no pré-processamento do texto do fórum. Como o conteúdo do fórum se encontra em linguagem natural, antes de se aplicar qualquer método estatístico de cálculo de relevância de termos, é feita uma triagem para separar os termos não representativos (*stopwords*), como advérbios, adjetivos, artigos e preposições, ou seja, termos que, em geral, não acrescentam representatividade ao documento e, muitas vezes, estão presentes somente para conectar frases.

As *stopword* são termos considerados não relevantes na análise dos textos, pois não traduzem sua essência. Sendo assim, antes que o arquivo do fórum se torne um documento no sistema, é feita uma verificação nas palavras contidas nesse arquivo e, logo, a eliminação das *stopwords*. Para possibilitar eliminação das *stopwords*, foi feito o uso de uma lista disponibilizada pela Google. No momento da criação do documento, foi feita uma verificação com vistas a averiguar se as palavras (termos) existentes no documento estavam contidas na lista de *stopwords*. Assim sendo, o documento criado possui somente os termos mais representativos, podendo, então, realizar o cálculo da sua representatividade.

3. Como já mencionado neste trabalho, a ponderação de termos (*Term Weighting*) é uma importante ferramenta para determinar a relevância de uma palavra, tanto em relação ao documento em que está contida, quanto em relação à coleção de documentos.

A principal abordagem utilizada para o cálculo de peso de um termo é o proposto por Salton (1988) TF*IDF. Onde TF é igual a:

$$TF(t_i, d_j) = \frac{f_{i,j}}{max_i f_{i,j}}$$

Onde o máximo é calculado sobre as frequências ***fi,j*** de todos os termos ***ti*** que ocorrem no documento ***dj***. E ***fi,j*** é a frequência do termo ***i*** no documento ***dj***.

A frequência inversa do documento (IDF) foi proposta inicialmente por Jones (1972). A ideia por trás desse algoritmo é de que um termo de consulta que ocorre em muitos documentos não é um bom representante, e, portanto, deve ter menor peso do que os que ocorrem em apenas alguns documentos. O cálculo baseia-se na contagem do número de documentos que contém o termo em questão na coleção que está sendo procurada.

$$idf = \frac{N}{n_i}$$

Onde ***N*** é igual ao número de documentos da coleção e $\boldsymbol{n_i}$ é igual a quantidade de documentos em que o termo ocorre pelo menos uma vez.

A essência do funcionamento do método TF*IDF consiste em determinar o quão relevante uma palavra é em relação a um conjunto de documentos. As palavras que são comuns em um pequeno grupo de documentos, ou em apenas um documento, tendem a ter pesos TF*IDF mais elevados do que as palavras comuns a todos os documentos. Isso se dá porque essa abordagem foi idealizada para se trabalhar com consultas (query), e uma palavra que está presente em todos os documentos não apresenta nenhuma representatividade, haja vista que uma vez que ela é utilizada na busca, todos os documentos serão recuperados, e o problema de se encontrar a informação relevante não seria resolvido.

A proposta de Salton e Buckley (1988), ao utilizar o IDF junto com o TF, foi a solução encontrada para poder reduzir o peso dos termos presentes em uma grande quantidade de documentos e privilegiar os termos que possuem uma alta frequência em um conjunto pequeno de documentos. Entretanto, o objetivo do uso da ponderação de termos neste trabalho é identificar os principais tópicos debatidos no fórum. Além do mais, com a extração das *stopwords* é possível eliminar grande parte dos termos que não apresentam significância ao contexto da discussão. Dessa maneira, para que se possa identificar as palavras-chave, faz-se necessário utilizar um algoritmo que atribua pesos mais significativos aos termos mais frequentes na coleção de documentos, ou seja, nos tópicos do fórum.

Pelo fato de dar menor peso aos termos que são frequentes em documentos, mas não tão frequente na coleção, o TF*IDF não se torna adequado para a resolução do problema em questão, pois, segundo Bun e Ishizuka (2002), este algoritmo tende a dar um peso mais significativo aos termos quando eles aparecem em apenas um documento, através da multiplicação do IDF.

Proposto em 2001 por Bun e Ishizuka, o TF*PDF (*Term Frequency * Proportional Document Frequency*) é uma abordagem que busca atribuir pesos mais significativos aos termos mais frequentes na coleção de documentos. Segundo Bun e Ishizuka (2002), esse método inovou de forma a dar mais peso aos termos que aparecem com frequência em muitos documentos. Em sua abordagem inicial, proposta por Bun e Ishizuka (2002), o algoritmo TF*PDF é usado para reconhecer os termos que explicam os principais

temas de cada arquivo de notícias (Hot Topics) semanais. Sua proposta baseia-se no conceito de que, sempre que houver um hot topic "no ar", o tema será discutido com frequência em muitos documentos e fontes de notícias. Os termos que explicam os *Hot Topics* que aparecem com frequência em muitos documentos serão ponderados de forma significativa.

Diferente da atribuição convencional de peso trabalhado no método TF*IDF, no algoritmo TF*PDF o peso de um termo é linearmente proporcional à frequência, e exponencialmente proporcional à relação do documento que contém o termo. Sendo assim, o algoritmo PDF é representado pela fórmula a seguir:

$$pdf = exp\frac{n_i}{N}$$

Desde sua proposta inicial, em 2001, o TF*PDF tem se apresentado como uma excelente ferramenta no que se refere à mineração de texto para a detecção de tópicos, tanto em um documento como em um conjunto de documentos. Diversos trabalhos, nos últimos anos, têm demonstrado a eficiência desse algoritmo, tais como: Jahnavi e Radhika (2012); Zhe et al. (2012); Barreira e Souza (2011); Ren et al. (2011); Kaur e Gupta (2012); Ma (2011), entre outros.

Sendo assim, este trabalho faz uso desse método como ferramenta de mineração de texto, com o propósito de extrair os principais temas debatidos nos fóruns de aprendizagem. Para se calcular a frequência do termo (TF), foi adotada uma abordagem diferente. Como não é realizado o cálculo de similaridade entre documentos da coleção, os termos são agrupados como um único documento, e os pesos TF's são calculados para todos os termos em relação a um único documento. Uma vez feito isso, o cálculo do peso PDF é realizado levando em consideração a frequência do termo em relação ao conjunto de documentos, a variável **ni** (número de ocorrência do termo na coleção de documentos, ou seja, nos tópicos do fórum).

Com o processo de ponderação de termos realizado através do algoritmo TF*PDF, os termos dos documentos são ranqueados com o objetivo de identificar as palavras que apresentam maior representatividade, ou seja, que supostamente trazem a temática da discussão do fórum.

4. Após a identificação dos principais tópicos discutidos no fórum, o sistema seleciona os seis primeiros termos e os submete a um motor de busca. No exemplo aqui proposto, foi utilizado a API do motor de busca Bing, que oferece

cinco mil consultas gratuitas por mês, e permite que os links sejam retornados tanto em XML quanto em JSON.

Uma interface de programação de aplicativos (API), como o fornecido pela Microsoft Bing Search, é um conjunto publicado de especificações para uma biblioteca de software host. O aplicativo pode se comunicar com a API para usar os serviços que a biblioteca oferece. A API Bing Search permite usar os dados que são recolhidos pelo motor de busca Bing em seu aplicativo de desktop, web, ou outro componente de software. Pode-se usar o Bing Search API com qualquer componente de software capaz de emitir uma solicitação HTTPS.

O Bing Search API pode obter resultados de pesquisa em seis categorias, descritas no quadro a seguir.

Quadro 1: Categorias dos resultados da API Bing Search

Categoria	Descrição
Web General	Páginas web que estão relacionadas a query.
Image	Imagens e ícones que estão relacionadas a query.
Video	Vídeos que estão relacionadas a query.
News	Notícias de artigos que estão relacionadas a query.
Related Search	Pesquisa expressões e seus correspondentes Bing Uniform Resource Identificadores (URIs) que estão relacionadas a query. Por exemplo: se a sua consulta é San Francisco, os Bing Search API retorna URIs que você pode usar para obter resultados de pesquisa sobre o Tempo em San Francisco, San Francisco Viagem, e San Francisco Coisas para Fazer. Você pode usar as expressões de pesquisa relacionados a sugerir consultas mais específicas para seus usuários.
Spelling Suggestions	Sugestões de ortografia para quaisquer palavras com erros ortográficos na sua consulta. Por exemplo: se a sua consulta é San Francsco bseball, a API do Bing retorna um único resultado com ambas as palavras com erros ortográficos corrigidos. Se a consulta não tem erros de ortografia, a API retorna um conjunto de resultados vazio. Você pode usar as sugestões de ortografia para sugerir consultas alternativas para os usuários.

Fonte: O autor.

Cada item recuperado contém informações sobre o recurso relevante, tal como o seu título, descrição e URI. O número e tipo de campos para cada resultado variam por categoria. No caso de resultados de imagem e vídeo, cada um deles contém informações miniatura que podem ser usadas para exibir as miniaturas na sua aplicação.

Além de usar as categorias de personalizar os resultados, pode-se dizer ao Bing Search API para filtrar os resultados em um conjunto. Por exemplo, pode-se limitar os resultados da pesquisa para um tipo de documento particular (por exemplo, Microsoft Word ou PDF), ou de uma região geográfica específica. É possível, ainda, especificar que as imagens devem ser de um determinado tamanho ou de uma determinada proporção, bem como solicitar apenas vídeos de certo período. Na categoria notícias, elas podem ser filtradas por entretenimento, política ou esportes.

A API Bing Search permite usar os parâmetros no quadro a seguir para especificar o número de resultados que se deseja obter; por onde começar no conjunto de resultados (ou seja, a paginação); e o formato resultado.

Quadro 2: Parâmetros da API bing search

Parâmetro	Descrição
$top	Especifica o numero de resultados retornados. Por padrão são retornados 50 para WEB, Imagens e Vídeos, e 15 para notícias.
Example: https://api.datamarket.azure.com/Bing/Search/Web?Query=%27Xbox%27&$top=10	
$skip	Especifica o ponto de partida dos resultados. O padrão é começar do zero.
Example: https://api.datamarket.azure.com/Bing/Search/Web?Query=%27Xbox%27&skip=20	
$format	Especifica o formato da resposta. As opções atuais são Atom (para XML) ou JSON. Padrão: Atom
Example: https://api.datamarket.azure.com/Bing/Search/Web?Query='Xbox'&$format=json	

Fonte: O autor.

A API de Bing Search permite que os resultados da busca sejam recebidos em dois diferentes formatos: JSON e XML.

Quadro 3: Formato retornados pela API bing search

Formato	Descrição
JavaScript Object Notation (JSON)	Um formato de intercâmbio de dados leve. JSON é apoiado por muitas linguagens de programação e ambientes, como o Microsoft Visual Basic, Java, Perl, PHP, Ruby, C ++ e C #.
Extensible Markup Language (XML)	Uma linguagem de marcação amplamente suportado que define um conjunto de regras para a codificação de documentos em um formato que seja legível e legível por máquina.

Fonte: O autor.

Os links retornados são apresentados contendo os seguintes elementos:

Quadro 4: Elementos que compõem os arquivos retornados pela API bing search

Nome	Descrição
ID	Identificados
Title	Texto especificado na tag HTML <title> da página retornada
Description	Texto de descrição da página
DisplayUrl	URL Web que será apresentada ao usuário.
Url	URL completa

Fonte: O autor.

5. Após feita a consulta, os links (no caso deste trabalho, somente os dez primeiros) são apresentados ao usuário como sugestão para leitura e aprofundamento no estudo.
6. Um mecanismo de avaliação é proposto com a finalidade de verificar a relevância do link para fórum para o qual foi proposto. O usuário poderá avaliar positivamente (gostei) ou negativamente (não gostei), de acordo com o enquadramento do documento em meio à discussão proposta no fórum.
7. Uma vez que o usuário gosta de um link recomendado, é feito um processo de indexação do campo descrição (que apresenta um resumo do documento) e guardam-se os termos de maior peso no perfil do usuário, para que, em um momento posterior, possam ser feitas sugestões de outras matérias, com base em seu interesse.

CAPÍTULO VI

A PESQUISA

Com o objetivo de verificar a eficiência da implementação do algoritmo apresentado neste trabalho, foi realizada uma pesquisa qualitativa em onze diferentes disciplinas, de três cursos técnicos oferecidos pela Universidade Estadual do Maranhão (UEMA), através do Núcleo de Tecnologias para Educação (UemaNet). As disciplinas escolhidas para a pesquisa foram: Análise de Sistemas; Projeto de Redes; Interação Homem Computador; Gerência de Projetos; Programação Orientada a Objetos; Linguagem de Programação II; Sistemas Operacionais; Fundamentos de Informática; Aplicações Web I; Aplicações Web II e Projeto e Desenvolvimento de Sistemas.

O UemaNet está posicionado na UEMA como o segmento responsável pela coordenação da modalidade de EAD e por outras ações educacionais que demandam a utilização de recursos tecnológicos. Vislumbra o atendimento às demandas da sociedade maranhense no que concerne à formação de profissionais nas diversas áreas do conhecimento em nível técnico, no âmbito da educação profissional, superior (graduação e pós-graduação) e formação continuada. O núcleo está subordinado diretamente à Reitoria da UEMA, e se articula com as Pró-Reitorias e Centros de Ciências e de Estudos Superiores, objetivando assegurar a integração de esforços e a otimização de recursos para o pleno desenvolvimento das suas ações.

A UemaNet presta suporte tecnológico à educação presencial e é responsável pela concepção, intermediação, gestão, avaliação e difusão de projetos educacionais na modalidade a distância da UEMA. Com a missão de "promover educação com qualidade e responsabilidade sócio-ambiental", o referido núcleo busca contribuir para a formação de cidadãos comprometidos com a sociedade e que tenham atos responsáveis, social e ambientalmente. A estrutura de gestão do UemaNet está organizada de forma descentralizada, sendo que, até o momento, atende um total de trinta e seis polos, organizados para fornecer suporte às atividades acadêmicas e encontros presenciais.

A UEMA, por meio do UemaNet, atua junto à Secretaria de Educação Profissional e Tecnológica (SETEC), vinculando-se ao Programa Escola Técnica Aberta do Brasil (Rede e-Tec Brasil) com a oferta de cursos técnicos de nível médio a distância. Todas as disciplinas escolhidas para a realização da pesquisa aqui apresentada fazem parte da Rede

e-Tec, que são os seguintes cursos: Técnico em Informática; Técnico em Redes de Computadores; e Técnico em Gestão em TI.

O objetivo desta pesquisa foi verificar a eficiência de palavras-chave dos tópicos de discussão do AVA na extração automática, bem como averiguar se o uso desses termos proporcionou a recuperação de links relevantes ao tema em discussão. Para essa averiguação, foram extraídos diversos tópicos (em média quinze) de cada disciplina, salvos em arquivos de texto (*.txt*) e submetidos à execução do algoritmo. O sistema, por sua vez, fazia a extração de palavras-chave, ordenava-as e as submetias a seis palavras de maior peso para a API do motor de busca Bing, que retornava os links (os dez primeiros), os quais seriam sugeridos. Feito isso, foi gerado um formulário contendo a descrição da proposta do fórum, as seis palavras-chave de maior relevância, segundo o algoritmo proposto, os links recomendados, e os tópicos utilizados para a execução dos testes, conforme se pode visualizar no Apêndice 1.

Para avaliar a eficiência, ninguém melhor do que os próprios professores que administravam as disciplinas. Dessa forma, foi-lhes encaminhado um formulário contendo os itens descritos acima e mais duas perguntas:

1. As palavras chaves extraídas possuem representatividade em relação ao texto dos fóruns?
   ( ) Discordo Completamente
   ( ) Discordo Parcialmente
   ( ) Indiferente
   ( ) Concordo Parcialmente
   ( ) Concordo Completamente
2. Os links sugeridos estão de acordo com o tema da discussão?
   ( ) Discordo Completamente
   ( ) Discordo Parcialmente
   ( ) Indiferente
   ( ) Concordo Parcialmente
   ( ) Concordo Completamente

Dos onze questionários elaborados e submetidos à avaliação dos professores, dez responderam que concordavam completamente em ambas as questões, sendo que apenas um respondeu que concordava parcialmente. Sendo assim, apesar da pequena amostra utilizada na pesquisa, pode-se verificar a eficiência da captura automática das palavras-chave e, consequentemente, a apresentação de links relevantes à temática debatida no fórum. Ao

atribuir maior relevância aos termos que apresentam maior frequência, o algoritmo TF*PDF consegue identificar os assuntos mais discutidos, ou seja, os *hot topics*.

Silva e Borba (2010) destacam que o fórum possibilita uma discussão aberta e de longa duração, o que pode se prolongar até a conclusão da disciplina. Por esse motivo, a escolha dos temas a serem discutidos deve possibilitar uma discussão mais ampla, em que a troca de experiências entre os participantes ocorra através da soma ou contraposição de ideias apresentadas. Sendo assim, ao elaborar um fórum de discussão, os professores costumam não delimitar ou restringir a temática em um assunto muito específico, para que os participantes possam correr a discussões por diferentes vertentes.

Por apresentar essa característica dinâmica, os fóruns de discussões dificultam a elaboração e seleção prévia de um material de leitura extra, que venha estimular as discussões no decorrer do processo de construção do conhecimento de forma colaborativa. Além do mais, algumas das principais características dos cursos de EAD são o planejamento e elaboração prévia de toda a sala de aula virtual. Dessa forma, a ferramenta de mineração de texto proposta neste trabalho apresenta-se como um poderoso recurso, ao oferecer uma análise dinâmica do conteúdo discutido nos fóruns para, então, poder sugerir fontes que estejam diretamente relacionadas com a temática debatida.

Apesar de o algoritmo ter se mostrado bastante eficiente na extração dos *hot topics* dos fóruns, através da implementação da pesquisa pode-se verificar algumas vulnerabilidades. Na disciplina em que o professor avaliou com “concordo parcialmente”, identificou-se que o tema principal da discussão era “Sistemas Operacionais”, entretanto, na maioria dos tópicos os participantes referenciavam-no apenas pela sigla “SO”. Dessa maneira, o processo remoção de *stopwords* acabava desconsiderando esse termo e, consequentemente, afetando o processo de extração.

Outro problema identificado nessa pesquisa está relacionado aos ruídos nos dados, ou seja, frequentemente são encontrados erros ortográficos nos textos analisados, como palavras sem acento, ausência de espaço entre as palavras, e outros.

Mais um problema percebido através desse experimento foi referente à flexão das palavras. Em mais de uma situação, o algoritmo capturou palavras no singular e também no plural, como, por exemplo: em um fórum de debate em que a temática abordava a arquitetura cliente/servidor, foi capturado o termo “cliente” e também o termo “clientes”. Para tratar esse problema, pretende-se, futuramente, incorporar a técnica de normalização morfológica (*stemming*), que consiste na eliminação das variações morfológicas de uma palavra através da

identificação do radical da mesma. Essa técnica de identificação de radicais é denominada lematização ou *stemming*, que em inglês significa reduzir uma palavra ao seu radical (ou raiz).

Desse modo, com a aplicação da normalização morfológica, resolve-se não somente o problema da flexão das palavras, mas também outros, como, por exemplo, a ambiguidade, que, apesar de não visualizada a sua ocorrência neste experimento, configura mais uma barreira nesse processo.

CONSIDERAÇÕES FINAIS

Como já apresentado neste trabalho, a interação com o conhecimento e com as pessoas é fundamental para o processo de aprendizagem. Sendo assim, os fóruns de discussão de um AVA buscam oferecer um meio pelo qual os seus participantes interajam de forma a transformar um determinado grupo de informações em conhecimento. As múltiplas interações e trocas comunicativas entre seus participantes possibilitam que esses conhecimentos sejam reconstruídos e reelaborados. Entretanto, essa participação exige preparo através de leituras adequadas, permitindo que haja a troca de opiniões pessoais fundamentadas, sem dar espaço ao "achismo".

O sistema proposto neste trabalho busca oferecer um meio automático de recomendar links para fomentar as discussões nos fóruns. O algoritmo TF*PDF proporciona uma forma de extrair os tópicos mais importantes, levando em consideração que eles aparecem com bastante frequência. Portanto, esse método difere do modelo clássico de ponderação de termo e pesos proposto por Salton e Buckley (1988), ao darem pesos mais significantes aos termos mais frequentes.

Diferentemente das suas abordagens iniciais, a utilização do algoritmo TF*PDF neste trabalho não fez utilização calculo de similaridade entre documentos, pois nesta aplicação foi delegada esta tarefa ao motor de busca reduzindo a complexidade do algoritmo.

Os testes preliminares do algoritmo aplicado a fóruns de discussão apontaram algumas lacunas que devem ser tratadas antes de se colocar a ferramenta em um ambiente de produção. Entende-se que esses testes devem ser estendidos para que se possa ampliar a observação do comportamento do algoritmo, e, talvez, encontrar outros problemas a serem tratados.

Futuramente, pretende-se incorporar o algoritmo implementado em um plugin para o Moodle, possibilitando testes mais ostensivos e a sua utilização em produção. Outra ideia é conectar o mecanismo de busca (metabusca) a um repositório de objetos de aprendizagem, dando mais garantias sobre a qualidade dos itens recuperados.

REFERÊNCIAS

ALAG, S.; MACMANUS, R. **Collective intelligence in action**. Manning New York: B&W, 2009.

ALMEIDA, Maria Elizabeth Bianconcini de. Educação a distância na Internet: abordagens e contribuições dos ambientes digitais de aprendizagem**. Educação e Pesquisa**, São Paulo, v. 29, n. 2, p. 327-340, jul./dez. 2003.

ALVES, Lucineia. Educação a distância: conceitos e história no Brasil e no mundo. **Associação Brasileira de Educação a Distância**, v. 10, 2011.

BARREIRA, Rafael Gonçalves; SOUZA, Jackson Gomes. Proposta de uma ferramenta de notificação de conteúdo do Twitter baseado em técnicas de Extração Automática de Tópicos e Clustering. In: ENCONTRO DE COMPUTAÇÃO E INFORMÁTICA DO TOCANTINS, 13, 2011, Palmas. **Anais...** Palmas: CEULP/ULBRA, 2011. p. 219-227.

BRASIL. Lei nº 9.394, de 20 de dezembro de 1996.

BUN, Khoo Khyou. ISHIZUKA, Mitsuru. Information Area Tracking and Changes Summarizing in WWW. **International Conf. on WWW and Internet**, Orlando, Florida, p. 680-685, 2001.

BUN, Khoo Khyou. ISHIZUKA, Mitsuru.Topic Extraction from News Archive Using TF*PDF Algorithm. Proceedings of the 3rd International Conference on Web Information Systems Engineering. p. 73-82, 2002.

COWIE, J.; LEHNERT, W. Information Extraction. **Communications of the ACM**, New York, v. 39, n. 1, jan. 1996

FARIA, Adriano A.; SALVADORI, Angela. A Educação A Distância E Seu Movimento Histórico no Brasil. **Revista das Faculdades Santa Cruz**, v. 8, n. 1, jan./jun. 2010.

FONSECA, Luis Carlos Costa. **Tese de Doutorado: Um Assistente Pessoal de Aprendizagem Continuada na Web.** UFRGS. 2009

GUAREZI, R. C. M; MATOS, M. M.. **Educação a distância sem segredos**. Curitiba: Ibpex, 2009.

HAN, J; KAMBER, M.. **Data Mining**: Concepts and Techniques. 2. ed. Morgan: Kaufmann, 2006.

JAHNAVI, Y. RADHIKA, Y. **A Cogitate Study on Text Mining**. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, v.1, Issue-6, aug. 2012.

JANNACH, Dietmar; ZANKER, Markus; FELFERNIG, Alexander; FRIEDRICH, Gerhard. **Recommender Systems:** An Introduction. New York: Cambridge, 2011.

JONES, Karen Spärck. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation.** v. 28, n. 1, p. 11-21, 1972.

KAUR, Kamaldeep; GUPTA, Vishal. A Survey of Topic Tracking Techniques. **International Journal of Advanced Research in Computer Science and Software Engineering.** v. 2, n. 5, mai. 2012

KENSKI, V. M.. Processos de interação e comunicação mediados pelas tecnologias. In: ROSA, D., SOUZA, V. (Orgs.). **Didática e práticas de ensino:** interfaces com diferentes saberes e lugares formativos. Rio de Janeiro: DP&A, 2002.

KONCHADY, Manu. **Text Mining Application Programming.** Boston: Thomson Leatning Inc. 2006

KUECHLER, W. L. Business applications of unstructured text. **Communications of ACM**, v. 50, n. 10, p. 86-93, 2007.

LOPES, M. C. S.. **Mineração de Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português.** PhD thesis. Universidade Federal do Rio de Janeiro, 2004.

MA, Hui-Fang. Hot Topic Extraction Using Time Window. **International Conference on Machine Learning and Cybernetics,** Guilin, p. 10-13, July, 2011.

MAIA, Carmem; MATTAR, João. **ABC da EaD.** São Paulo: Pearson, 2007.

MOOERS; Calvin N. Zatacoding applied to mechanical organization of Knowled. **American Documentation**, v.2, p. 20-32, 1951.

MOORE, Michael. KEARSLEY, Greg. **Educação a Distância; uma visão integrada.** São Paulo: Thomson Learning, 2007.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L.. **Mineração de Textos.** Instituto de Informática. Universidade Federal de Goiás, 2007.

MUGNOL, Marcio. A Educação a Distância no Brasil: conceitos e fundamentos**. Rev. Diálogo Educ**., Curitiba, v. 9, n. 27, p. 335-349, maio/ago. 2009.

NUNES, I. B.. A história da EAD no mundo. In: LITTO, F. M. e FORMIGA, M. (Orgs.) **Educação a distância o estado da arte**. São Paulo: Pearson Education, 2009.

OLIVEIRA, Gerson P. O Fórum em um Ambiente Virtual de Aprendizado Colaborativo. **Revista Digital de Tecnologia Educacional e Educação a Distância**. São Paulo, v. 2, n. 1, 2005.

PETERS, Otto. **Educação a Distância em Trasição.** Germani: Unisinos, 2004.

PIVA, D. PUPO, R. GAMEZ, L. OLIVEIRA, S. **EaD na Prática.** Planejamento, métodos e ambientes de educação online. Rio de Janeiro: Elsevier, 2011.

PRETI, O.. **Educação a Distância**: uma prática educativa mediadora e mediatizada. Cuiabá: NEAD/ IE –UFMT, 1996.

REN, Yuyan; DU, Yajun; HUANG, Xiaoping; XU, Yong. Topic Detection of News Stories with Formal Concept Analysis. **Journal of Information & Computational Science**. v. 8, n. 9, p. 1675-1682, 2011.

REZENDE, Solange O.; MARCACINI, Ricardo M.; MOURA, Maria F.. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. **Revista de Sistemas de Informação da FSMA**, n. 7, p. 7-21, 2011.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha; KANTOR, Paul B.. **Recommender Systems Handbook**. New York: Springer, 2011.

ROMANI, Luciana Alvim Santos; ROCHA, Heloísa Vieira da; SILVA, Celmar Guimarães. Ambientes para educação a distância baseados na Web: Onde estão as pessoas ? In: Workshop sobre fatores humanos em sistemas computacionais, 3, 2000, Gramado. **Anais...** Gramado: UFRGS, 2000. p.12-21.

SALTON, G; MCGRILL, M. J. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill Book Co., 1983.

SALTON, Gerard; BUCKLEY, Christopher. Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, v. 24, n. 5, p. 513-523, 1988.

SCARINCI, Rui Gureghian. **SES: Sistema de Extração Semântica de informações.** Dissertação de mestrado. Instituto de Informática. Universidade Federal do Rio Grande do Sul, 1997.

SILVA, Marco (Org.). **Educação online**. São Paulo: Loyola, 2006.

SILVA, Rose Madalena Pereira da.; BORBA, Sara Ingrid. Fórum de Discussão Como Ferramenta para a Construção do Conhecimento. **V Encontro de Pesquisa em Educação de Alagoas (EPEAL).** ISSN 1981-3031. 2010.

WIVES, Leandro K. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. Exame de Qualificação. Universidade Federal do Rio Grande do Sul, , Rio Grande do Sul, 2002. Programa de Pós-Graduação em Computação.

YATES, Ricardo Baeza; RIBEIRO NETO, Berthier Ribeiro. **Modern Information Retrieval**: the concepts and technology behind search. New York. 2 ed. Addison-Wesley. 2011.

YATES, Ricardo Baeza; RIBEIRO NETO, Berthier. **Modern Information Retrieval**. New York: ACM Press, 1999.

ZHE, Gong et al. An Online Hot Topics Detection Approach Using the Improved Ant Colony Text Clustering Algorithm. **Journal of Convergence Information Technology(JCIT**), v.7, n. 2, february 2012.