Recuperação de Informação sobre Currículos com Geolocalização: um protótipo usando busca vetorial

Amarildo Martins de Magalhães
Instituto Federal de Minas Gerais - IFMG, Email: amarildo@gmail.com

Renata Maria Abrantes Baracho
Universidade Federal de Minas Gerais - UFMG, Email: renatabaracho@ufmg.br

RESUMO

O Instituto Federal Minas Gerais (IFMG) recebeu do governo federal R$ 790.575,82 em 2013 e R$ 1.109.576,55 em 2014 para auxílio a pesquisadores e incentivo a pesquisa. Todo esse investimento requer que os programas de pós-graduação sejam acompanhados tanto do ponto de vista regulatório, como de gestão, e para isso, ferramentas que possibilitem melhorar a recuperação da informação sobre pesquisadores e pesquisa tornam-se importantes. Considerando a inexistência de tais ferramentas no IFMG, a questão dessa pesquisa aborda a melhoria dos processos de recuperação da informação sobre currículos Lattes de profissionais e pesquisadores. Por meio de uma pesquisa aplicada, foi desenvolvido um protótipo usando busca vetorial nos currículos Lattes do IFMG e criado um ranking com geolocalização, haja vista que os servidores estão espalhados em 17 cidades. Ao final, o protótipo é aplicado e avaliado sob uma perspectiva qualitativa pela equipe de gestão de pesquisa do IFMG. Os resultados mostram a validade do protótipo e que a geolocalização pode ser essencial para a criação de um mapa do conhecimento organizacional.

Palavras-chave: Recuperação da Informação. Modelo Vetorial de Recuperação da Informação. Lattes. Currículos. IFMG. Geolocalização. Scrapy. Crawler.

---

1 INTRODUÇÃO

A informação como objeto de estudo, constitui insumo para uma ciência inevitavelmente interdisciplinar, dada sua utilização invariável em todas as áreas do conhecimento. O volume de informação desestruturada disponível na era de *Big Data*, promove a necessidade de melhorar e aplicar técnicas de Recuperação da Informação (RI) que possibilitem aos usuários uma profundidade e personalização maior em suas buscas. Nesse contexto, torna-se importante a apresentação de resultados considerando a geolocalização do usuário, conforme modelo utilizado pelo Google em suas buscas. Logo, além de precisão e pertinência, existem fatores essenciais a serem considerados nos modelos de RI como a semântica, imprecisão e localização.

Nos últimos anos houve um valor considerável investido em pesquisa nas instituições federais de ensino. O Instituto Federal Minas Gerais (IFMG) recebeu do governo federal R$ 790.575,82 em 2013 e R$ 1.109.576,55 em 2014 para auxílio a pesquisadores e incentivo a

pesquisa. A Coordenação de Pessoal de Nível Superior (CAPES) realiza processos de monitoramento e auditoria nos programas de pós-graduação oferecidos no Brasil e frequentemente é necessário que os gestores de pesquisa das instituições repassem ao MEC e a CAPES indicadores de produção científica. Embora exista uma variedade de bases de dados científicas, a base de currículos de profissionais e pesquisadores Lattes oferece informações integradas e é a única a fornecer informações importantes como participação em eventos, bancas e orientações (MAGALHÃES; QUONIAM; MENA-CHALCO; SANTOS, 2014). O Lattes pode ser utilizado como base para criação de indicadores de gestão ou acompanhamento da pesquisa, entretanto, a estrutura de RI da plataforma não permite uma personalização da busca pelos usuários e gestores institucionais.

A questão dessa pesquisa consolida-se em como melhorar a recuperação da informação sobre competências profissionais e acadêmicas dos servidores do IFMG? É necessário que os programas de pós-graduação sejam acompanhados tanto do ponto de vista regulatório, como de gestão e para isso ferramentas que possibilitem melhorar a RI de pesquisadores e pesquisa tornam-se importantes. Uma gestão mais eficiente dos processos de pós-graduação permite acompanhar se os valores investidos pelo governo geram retorno efetivo à comunidade científica e sociedade.

Esse trabalho, por meio de uma pesquisa aplicada, apresenta o desenvolvimento e aplicação de um protótipo baseado em busca vetorial para recuperação da informação de currículos dos servidores do IFMG, objetivando a melhoria desse processo. Como forma de aperfeiçoar a experiência do usuário nessa busca, o artigo apresenta a busca de currículos considerando geolocalização. Um objetivo específico da pesquisa é a criação de um mapeamento do conhecimento no IFMG e incentivo a criação de grupos locais de pesquisas similares. Para validação do protótipo desenvolveu-se uma pesquisa qualitativa junto a equipe da Pró-Reitoria de Pesquisa e Pós-Graduação do IFMG em uma abordagem qualitativa.

2 FUNDAMENTAÇÃO TEÓRICA

A informação como objeto de estudo, constitui insumo para uma ciência inevitavelmente interdisciplinar, dada a utilização invariável desse insumo em todas as áreas do conhecimento. Há uma recorrente tentativa de se fundamentar essa ciência sobre alguns pilares, como a

Biblioteconomia e Ciência da Computação. No entanto, conforme Wersig (1993) é necessária uma perspectiva pragmática e evolucionária que constitua as interfaces entre diferentes áreas como subsídios para reconstrução permanente da área. Sobretudo, o direcionamento dessa ciência se deve à conexão entre pessoas, informação e tecnologia, conforme descrito na missão das *I-Schools* (Escolas de Informação) espalhadas pelo mundo (SEADLE, 2007).

Os diferenciais competitivos e de inovação necessários ao desenvolvimento e evolução de entidades públicas e privadas passam diretamente pela capacidade de adquirir, representar, processar e principalmente recuperar informação. O potencial de inovação e desenvolvimento das organizações está intrinsecamente ligado ao conhecimento existente em seus colaboradores e conforme aponta Seadle (2007), a informação é o conhecimento relevante para uma ação. Esse conhecimento pode ser classificado como tácito (residente na mente das pessoas, difícil de ser estruturado e transmitido) ou explícito (formalizado, documentos, arquivos, processos) que está presente nas pessoas, processos e rotinas de uma empresa (CHOO, 2006). Sendo vasto o acervo de informações disponíveis, e principalmente sua natureza não estruturada, torna-se imprescindível a utilização de processos eficientes de arquitetura da informação (SOUZA; ALMEIDA; BARACHO, 2013).

1.1 ARQUITETURA DA INFORMAÇÃO

A arquitetura da informação precisa permear três esferas (usuários, contexto e conteúdo), que formam a Ecologia Informacional de cada instituição. O contexto refere-se ao entendimento das metas institucionais e dos recursos disponíveis, o conteúdo refere-se ao volume de dados existente e os usuários referem-se ao entendimento dos comportamentos de pesquisa e necessidades informacionais (MORVILLE; ROSENFELD, 1998). A arquitetura da informação possui duas vertentes que contemplam diferentes fases do processo manipulação da informação. Em uma, destaca-se todo o fluxo abstrato existente na interação usuário, contexto e conteúdo, assim como, aspectos culturais, éticos e sociais na manipulação e uso da informação. Em outra, constam os processos de Representação do Conhecimento - *Knowledge Organization Systems* (KOS), técnicas de Recuperação da Informação e Gestão do Conhecimento. Os processos de recuperação da informação são o elo entre usuários e conteúdo e constituem a apresentação do resultado gerado por todas as etapas anteriores do processo de arquitetura.

1.2 RECUPERAÇÃO DA INFORMAÇÃO

Modelos de recuperação da informação quantitativos como booleanos, vetorial, probabilísticos ou mesmo os baseados em conjuntos fuzzy podem não satisfazer na plenitude as necessidades dos usuários. O sistema de informação está completo quando o usuário está satisfeito com o resultado da busca (BARACHO; CENDÓN, 2008) Atualmente, além de precisão e pertinência, existem fatores essenciais a serem considerados nos modelos de RI como a semântica, imprecisão e a localização. Na era do *Big Data*, é importante considerar recomendações baseadas em geolocalização. Por exemplo, um usuário procurando por um hotel específico, talvez possa se beneficiar de recomendações de bons restaurantes ou atrações próximas a aquele hotel.

1.3 PLATAFORMA LATTES[1]

A plataforma Lattes constitui uma importante forma de registrar o conhecimento por meio da produção científica e de profissionais e pesquisadores. Essa base de dados é atualmente no maior banco de currículos de pesquisadores do país, com mais de 3 milhões de currículos (BRITO; QUONIAM; MENA-CHALCO, 2016). As bases de periódicos, como Web of Sciense (ISI), Sielo, Scopus e Pubmed não possibilitam uma análise interdisciplinar sobre determinado assunto, haja vista as diferenças estruturais e ausência de integração (MAGALHÃES; QUONIAM; MENA-CHALCO; SANTOS, 2014). Existem informações integradas que estão disponíveis apenas nos currículos dos pesquisadores que não constam em bases de dados, como as citadas, como por exemplo, os projetos submetidos, patentes, orientações em andamento, dentre as outras produções (FERRAZ; QUONIAM, 2013). Uma base que tem se aproximado do conceito de base integrada e possibilitando a construção de mapas do conhecimento é a plataforma Lattes. No Brasil, os pesquisadores podem e precisam registrar suas informações profissionais, acadêmicas e de pesquisa no Lattes, mantido pelo Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq). A CAPES realiza processos de monitoramento e auditoria nos programas de pós-graduação oferecidos no Brasil. Para isso é necessário que os gestores de pesquisa repassem ao MEC e a própria CAPES indicadores de produção científica. É necessário compilar de maneira organizada as produções acadêmicas e

[1] http://lattes.cnpq.br/, Acesso em: 05 jun 2016

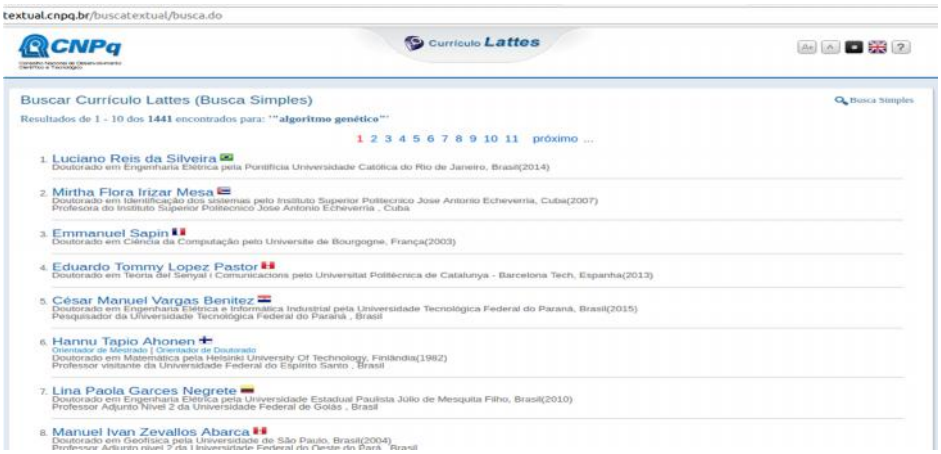
tecnológicas dos programas *Strictu Sensu* no país para a prestação de contas junto aos órgãos de avaliação do ensino superior conforme apontam (MARTINS; MACCARI; STOROPOLI; RICCIO, 2012).

No Lattes, os dados são postados pelo próprio pesquisador e não há validação destes, permitindo a inserção de dados não comprovados. Apesar disso, não se pode desconsiderar seu valor como base integrada de informações acadêmicas e profissionais (BRITO; QUONIAM; MENA-CHALCO, 2016). Se por um lado, a plataforma Lattes se destaca como uma fonte rica em informações, por outro a forma tradicional como estas informações podem ser acessadas e utilizadas pode restringir o seu uso (MAGALHÃES; QUONIAM; MENA-CHALCO; SANTOS, 2014).

O Lattes oferece a todos usuários uma busca por palavra(s) chave(s)[2] com o objetivo de fazer o mapeamento da produção científica por tema ou área. Nessa busca é possível especificar filtros como currículos apenas de doutores, nacionalidade e em que local do currículo os termos devem buscados. É possível também utilizar os operadores "AND", "OR", "NOT", "NEAR" para aperfeiçoar a relação entre os termos. As opções de filtro e busca para recuperação das informações são razoáveis, entretanto, o formato como os currículos são apresentados aos usuários dificulta o entendimento. O sistema retorna apenas a lista de currículos que contém determinadas palavra(s) chave(s), mas não exibe em que parte do currículo aquelas palavras chaves foram encontradas, não exibe quantificadores para as diferentes áreas do currículo, e principalmente a ordenação dos resultados não considera relevância ou geolocalização. Os resultados não são apresentados em ordem alfabética, não é possível identificar como o Lattes classifica os resultados apresentados. É relevante considerar a geolocalização dos currículos, dada a própria amplitude da base, que possui currículos não somente de pesquisadores nacionais. A Figura 1 apresenta o retorno do sistema para a busca "algoritmo genético":

[2] Disponível no endereço: http://buscatextual.cnpq.br/buscatextual/busca.do?metodo=apresentar

Figura 1 - Interface de apresentação de resultados do Lattes



Fonte: http://lattes.cnpq.br/, acesso em 17/junho/2016.

Se do ponto de vista da busca disponível a todos, há limitações no sistema de recuperação, do ponto de vista dos gestores dos programas de pós-graduação, as limitações são maiores. O sistema é disponibilizado pela CAPES aos gestores do endereço http://efomento.cnpq.br/efomento/ e permite que sejam realizados filtros, como por exemplo, "todas as publicações feitas no ano de 2016" para exportação de dados em formato "CSV". Essa versão não possibilita buscar currículos sobre determinados temas. A Figura 2 exibe a tela disponível para extração dos dados:

Figura 2 - Interface disponível aos gestores de programas de pós-graduação

Fonte: http://efomento.cnpq.br/efomento/, acesso em 17/junho/2016.

Os recursos disponíveis para recuperação da informação geram dificuldades de análise e compilação de informações sobre pesquisadores e pesquisas para auditoria dos programas de pós-graduação. Essas limitações geram um impacto tanto para auditoria, quanto para gestão, tendo em vista, a dificuldade em criar indicadores para gestão dos programas de pós-graduação.

1.4 EXTRAÇÃO DE DADOS DA PLATAFORMA LATTES

Dada as limitações existentes no sistema de recuperação do Lattes, existem uma quantidade considerável de trabalhos que realizam a extração de dados da base do Lattes. Os trabalhos realizados por Magalhães, Quoniam, Mena-Chalco e Santos (2014), Ferraz (2013) e Quoniam e Brito (2013), Quoniam e Mena-Chalco (2016) utilizam a ferramenta livre Scriptlattes para realizar esse processo. Essa ferramenta realmente apresenta a extração de informações acadêmicas e profissionais e permite também gerar gráficos de colaboração, de internacionalização da pesquisa e mapas geográficos de investigação dos pesquisadores. A ferramenta utiliza um *crawler*[3] escrito em linguagem de programação Python para extrair os currículos a partir da URL[4] de cada currículo.

Recentemente houve uma mudança no sistema do Lattes que impossibilitou o uso dessa ferramenta. O endereço de acesso aos currículos começou a exigir a entrada de CAPTCHA, que é um conjunto de alfanumérico para identificar se o visitante é um robô (*crawler*). No site do http://scriptlattes.sourceforge.net/ é possível visualizar uma petição dos autores do programa para retirada do CAPTCHA dos currículos Lattes. Há uma possibilidade de extração semiautomática, na qual, os currículos devem ser baixados um a um para uma máquina. É um processo que precisa ser avaliado considerando principalmente a extração de uma grande quantidade de registros.

1.5 PÓS-GRADUAÇÃO NO IFMG

Os Institutos Federais de Ciência, Tecnologia e Educação, autarquias formadas pela união das antigas Escolas Agrotécnicas e Escolas tecnológicas como o CEFET, recebem do governo

[3] Crawler é um programa que automaticamente recupera páginas web para criar um índice ou uma base local de páginas (GARCIA-MOLINA, 1999).
[4] URL – Localizador Padrão de Recursos ou Endereço digital de uma página

federal valores para desenvolvimento dos pesquisadores e iniciativas de pesquisa. O Instituto Federal de Ciência, Educação e Tecnologia de Minas Gerais (IFMG), autarquia formada pela união dos antigos CEFET-Ouro Preto, Escola Agrotécnica de Bambuí e Escola Agrotécnica de São João Evangelista, possui atualmente 1624 servidores (técnicos administrativos e professores), cerca de 12 mil alunos (níveis superior e técnico) e 17 câmpus. Nos últimos anos houve um investimento considerável na pesquisa da rede de educação federal profissional. A Tabela 1 apresenta um comparativo entre os valores gastos em Auxílio a Pesquisadores entre a Universidade Federal de Minas Gerais (UFMG) e o IFMG:

Tabela 1 - Comparativo gastos em Auxílio a Pesquisadores UFMG e IFMG

Instituição	Classificação Despesa	Ano	Valor
UFMG	Outras Despesas Correntes - Auxílio Financeiro a Pesquisadores	2013	R$ 359.879,53
UFMG	Outras Despesas Correntes - Auxílio Financeiro a Pesquisadores	2014	R$ 329.862,19
IFMG	Investimentos e Outras Despesas - Auxílio Financeiro a Pesquisadores	2013	R$ 790.575,82
IFMG	Investimentos e Outras Despesas - Auxílio Financeiro a Pesquisadores	2014	R$ 1.109.576,55

Fonte: http://www.portaldatransparencia.gov.br/, acesso em 18/junho/2016.

Analisando esses valores, há uma certa desproporcionalidade entre os valores gastos na UFMG e no IFMG, considerando principalmente o volume de produção científica, entretanto, não é escopo desse trabalho se aprofundar na análise desses dados, principalmente porque podem se tratar de classificações diferentes entre a gestão de cada instituição. A Tabela 1 demonstra o investimento e reforça a necessidade de auditar os programas de pós-graduação e projetos de pesquisa.

2 DESENVOLVIMENTO

Esse trabalho utiliza um modelo de busca vetorial com geolocalização para criar um protótipo de recuperação da informação sobre currículos dos servidores do IFMG.

2.1 METODOLOGIA

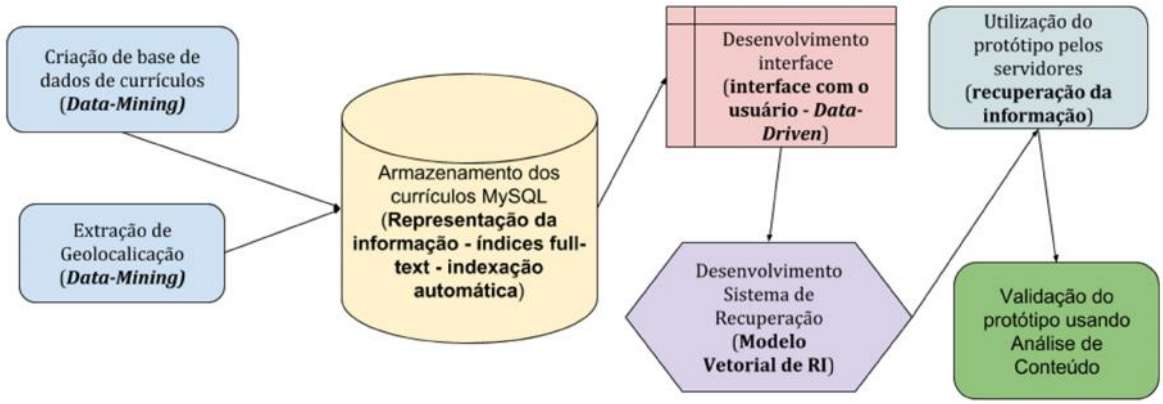
Utilizando os conceitos de Lakatos e Marconi (1991) e Gil (2009), a presente pesquisa pode ser classificada, quanto a natureza, como pesquisa aplicada, por gerar um resultado de aplicação

prática nas organizações, no caso, o IFMG e quanto a abordagem do problema, como pesquisa qualitativa, visto que os resultados e conclusões são gerados a partir de uma avaliação qualitativa. Dentro do campo da Ciência da Informação, esse trabalho pode ser classificado na área de Recuperação da Informação, *Data Mining* ou Descoberta de Conhecimento em Bases de Dados.

Considerando a ausência de uma base de dados de currículos dos servidores do IFMG, a primeira tarefa foi criar essa base. Em seguida, para recuperar a geolocalização, foi utilizada a API[5] do Google Maps de geolocalização. Os dados extraídos e processados foram armazenados no banco de dados MySQL. Após a importação dos dados, desenvolveu-se a interface do protótipo e em seguida o sistema de recuperação da informação (RI) utilizando o modelo de busca vetorial.

Com a base de dados estabelecida, a geolocalização e o sistema de RI, o protótipo foi apresentado a equipe da Pró-reitoria de Pesquisa, Inovação e Pós-Graduação (PRPPG) do IFMG para execução de testes. Os resultados foram compilados qualitativamente de acordo com essa validação. A Figura 3 exibe o processo de desenvolvimento da pesquisa e sua conexão com a área de ciência da informação em negrito:

Figura 3 – Processo de desenvolvimento do trabalho e sua conexão com a Ciência da Informação



Fonte: Elaborado pelo autor.

2.2 EXTRAÇÃO DE DADOS DE CURRÍCULOS DE PROFISSIONAIS

O CNPq limita o acesso automático aos currículos Lattes, mas existe uma ferramenta na Internet que consegue realizar essa extração automaticamente, o Escavador

[5] API - Interface de Programação de Aplicativo – permite integração entre sistemas

(www.escavador.com.br). As informações são disponibilizadas na Internet, mas não é possível identificar a metodologia usada por esse sistema para sincronização dos currículos do Lattes. Os dados apresentados são atualizados frequentemente, pois, foram avaliados manualmente alguns currículos de servidores do IFMG no Escavador e os mesmos estavam atualizados de acordo com a última versão do Lattes. Após identificar uma base com as informações necessárias dos currículos, foi necessário criar um *crawler* para extrair esses dados do Escavador. Existem diversas formas de implementar um *crawler*, dentre elas, uma muito utilizada é o Scrapy (http://scrapy.org/), que conforme aponta Yadav e Goyal (2015), é um dos *crawlers* mais robustos e utilizados para extração de dados da Internet.

O *Scrapy* (robô) foi configurado para permitir a extração das seguintes seções de cada currículo: *Nome do proprietário do currículo, resumo, formação, formação complementar, idiomas, áreas de atuação, organização de eventos, participação em eventos, participação em bancas, comissão julgadora de bancas, orientação, produções, outras produções, endereço profissional, atuação profissional, projetos e pesquisas em desenvolvimento e prêmios.* O robô foi configurado em dois níveis para fazer a extração dos currículos identificando cada servidor pelo nome. Como a extração dos dados se daria a partir do nome do servidor, foi criado um arquivo *JavaScript Object Notation (*JSON) a partir do banco de dados do ERP do IFMG, com endereço e nome dos servidores ativos, incluindo técnicos administrativos e professores. A lista continha um total de 1624 servidores em ordem alfabética.

O *Scrapy* foi executado para recuperar os currículos do escavador para os 1624 servidores do IFMG. A execução demorou cerca de 4 horas e retornou 2301 currículos. Como a extração dos dados é feita pelo nome do servidor, houve extração de mais de um currículo na existência de homônimo. Notou-se também que, para algumas pessoas, o Escavador possui mais de um currículo para o mesmo profissional, separando informações de seções distintas que deveriam estar no mesmo currículo. Exemplo: Nomes que possuem acentos existiam duas vezes na base do Escavador (uma com acento e outra sem).

Para melhorar a qualidade dos dados extraídos, foram realizados processos de comparação dos nomes de servidores com currículos repetidos. Por meio de um algoritmo criado pelo autor, foi feita a junção de servidores com nome e resumo do currículo iguais. O resultado da junção dos dados melhorou a base de dados, eliminando informações repetidas. No caso dos

currículos com homônimos não houve uma melhoria muito significativa. Ao final do processamento dos dados, foram compilados 1952 currículos aptos a serem utilizados. O fato do número final de currículos ser superior ao de servidores se deve a impossibilidade de junção de alguns currículos com nomes ou resumos diferentes, aos quais não se pode identificar a unicidade de informações. Apesar de alguns problemas, após uma avaliação junto a equipe da PRPPG, identificou que o resultado representa quase na totalidade os currículos dos servidores do IFMG, e decidiu-se pela utilização da base.

2.3 IDENTIFICAÇÃO DE COORDENADAS GEOGRÁFICAS

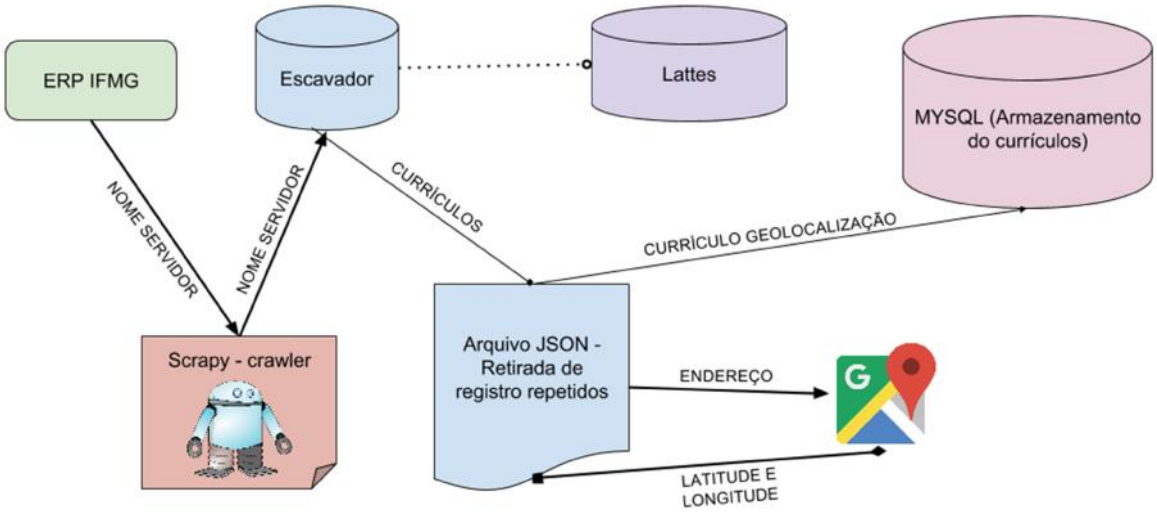
A ideia da geolocalização é permitir que os usuários do sistema identifiquem mais próximo de sua localização, quem está realizando pesquisas sobre determinado tema, assim como permitir identificar potenciais servidores com determinada competência. O endereço de cada servidor está disponível no seu cadastro dentro do ERP do IFMG, entretanto não há informação sobre latitude e longitude. Para isso, foi utilizada uma API do Google Maps para recuperar as coordenadas geográficas a partir do endereço. Foi desenvolvido um algoritmo em linguagem *Hypertext Preprocessor* (PHP) para recuperar as informações de geolocalização dos 1624 servidores a partir dessa API. Destes, apenas 6 estavam sem endereço no ERP e não foi possível encontrar as coordenadas. Após a execução do algoritmo, o sistema recuperou latitude e longitude de todos os 1618 servidores.

2.4 REPRESENTAÇÃO E ARMAZENAMENTO DOS DADOS

Inicialmente foi cogitada a possibilidade de armazenar os dados em um banco de dados não relacional, como o MongoDB, por sua facilidade em gerenciar índices para pesquisa em qualquer parte do documento. Entretanto, devido ao relacionamento bem definido das tabelas, optou-se pela utilização do banco relacional MySQL, principalmente por sua funcionalidade de criar índices FULLTEXT, que conforme aponta o manual do MySQL[6], permitem a implementação de um modelo de recuperação da informação vetorial. A Figura 4 exibe o processo de extração dos dados, processamento e geolocalização:

[6] Manual do MYSQL sobre índice FULLTEXT - disponível em: http://dev.mysql.com/doc/refman/5.7/en/fulltext-search.html, acesso em 18/junho/2016.

Figura 4 – Processo de extração dos dados, processamento e geolocalização

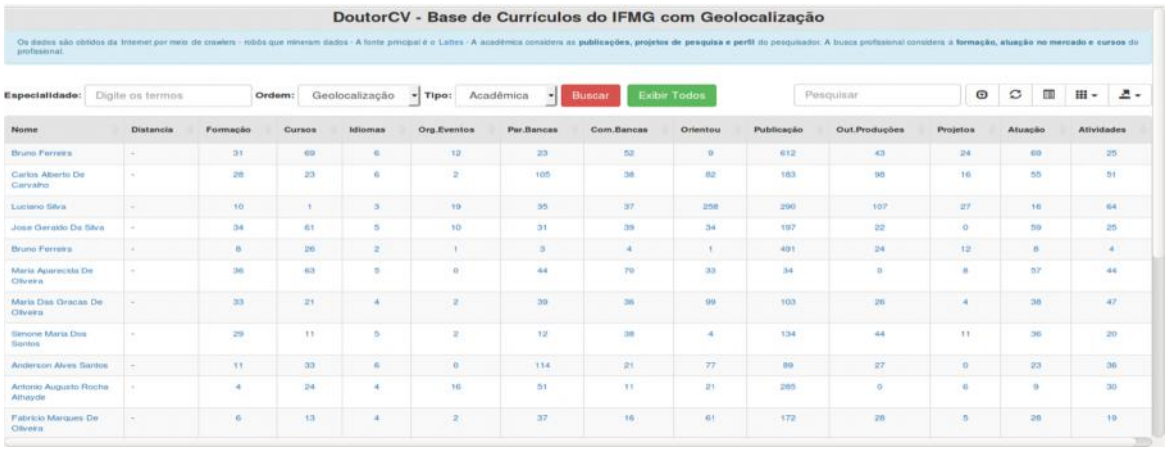


Fonte: Elaborado pelo autor.

2.5 DESENVOLVIMENTO DO PROTÓTIPO

O protótipo foi desenvolvido para ser utilizado em qualquer dispositivo, seja computador, tablet ou celular. Para isso, utilizou o framework Bootstrap que gerou páginas responsivas, ou seja, que se adaptam quando o usuário está acessando a partir de dispositivos móveis. Para desenvolvimento do protótipo, foi utilizado a linguagem de programação PHP com o banco de dados MySQL. O protótipo foi nomeado DoutorCV e foi disponibilizado no endereço www.doutorcv.com.br. Por critérios de segurança, os endereços dos profissionais foram ocultados do resultado da busca. As buscas são feitas e os usuários podem exibir os currículos mais próximos de sua localização. Para realizar esse cálculo, o protótipo pega a posição geográfica do usuário por meio do seu navegador. A Figura 5 exibe a interface do protótipo:

Figura 5 - Interface do protótipo



Fonte: Elaborado pelo autor.

2.5.1 Indicadores criados pelo protótipo

O protótipo exibe o nome do servidor na primeira coluna e na segunda coluna exibe a distância entre a localização atual do usuário e do pesquisador em quilômetros. Nas outras colunas, o protótipo exibe indicadores para cada parte do currículo, ou seja, se um servidor já publicou 3 artigos, o protótipo apresenta o número 3 na coluna Produções. O protótipo apresenta indicadores para as seguintes partes do currículo: *formação, formação complementar, idiomas, áreas de atuação, organização de eventos, participação em eventos, participação em bancas, comissão julgadora de bancas, orientação, produções, outras produções, endereço profissional, atuação profissional, projetos e pesquisas em desenvolvimento e prêmios.* O usuário pode visualizar os dados referentes a cada indicador apenas clicando sobre o número relativo. O protótipo exibe em uma nova janela, uma tabela com todos os dados daquela indicador. Nessa nova tabela, também é possível realizar buscas. Em sua tela inicial, o protótipo apresenta os servidores classificados por um somatório das seções de cada currículo. O sistema soma os quantitativos de cada área do currículo e classifica em ordem decrescente, ou seja, os primeiros servidores são os mais ativos no IFMG. Esse formato de apresentação dos dados visa criar um indicador quantitativo que permita aos gestores da PRPPG visualizar os servidores mais ativos sobre determinado tema.

2.6 CRIAÇÃO DE MODELO VETORIAL DE RI

Nesse trabalho, foi utilizado a busca vetorial do MySQL, através dos índices FULLTEXT. O modelo de busca desenvolvido utiliza a remoção de *stop-words*, em que artigos, preposições e conjunções são retirados dos termos digitados pelo usuário. Palavras consideradas não informativas, como (o, a, de, da, em), também são conhecidas como *stop-words* e foram ignoradas (SINGHAL, 2001). Outra técnica utilizada é o conceito de case-isensitive, que realiza a busca tanto em palavras maiúsculas quanto minúsculas, também são desconsiderados acentos. A busca vetorial é baseada em TF-IDF - frequência do termo inversa. No TF-IDF é levado em consideração a frequência do termo em uma ou mais colunas do documento (SINGHAL, 2001). O TF-IDF foi aplicado no protótipo através das funções do banco de dados MySQL MATH e AGAINST utilizando índices FULLTEXT em diversas colunas que representam dados do currículo como por exemplo, publicação, resumo, pesquisas em desenvolvimento, formação. O usuário pode realizar dois tipos de busca:

Busca vetorial:

Na busca vetorial tradicional, chamada *Natural Language Full-Text Searches*, o protótipo interpreta os termos de busca em linguagem natural (uma frase em texto simples). O protótipo utiliza a Fórmula 1 abaixo para realizar a busca:

Fórmula 1 – Busca vetorial por frequência do termo

$$\mathrm{w} = \mathrm{tf} * \mathrm{idf}$$

Na Fórmula 1, **w** é o peso, que é calculado com base na frequência do termo **tf** multiplicada pela frequência inversa **idf** do termo no currículo, ou seja, ele aumenta o peso de acordo com a quantidade de vezes que um termo aparece em um currículo, mas o diminui de acordo com o número de currículos que contém aquele termo. Por exemplo: Se “genético” aparece 5 vezes em um currículo, o peso é aumentado, mas se “genético” aparece em outros 1000 currículos, o peso é diminuído. O operador padrão da busca é “OU”.

Busca vetorial booleana em índices FULLTEXT:

Nesse tipo de busca, é possível especificar operadores para manipular a prioridade de determinados termos. Se algum termo obrigatoriamente precisa aparecer ou ser desconsiderado, o uso dos operadores + e - indicam isso respectivamente. Nesse trabalho, foi implementado esse

modelo para permitir a busca de currículos por expressões compostas. Por exemplo, a necessidade de recuperar apenas os currículos que contenham a expressão “Algoritmo Genético”, nesse caso, o protótipo busca apenas currículos que contenham as duas palavras juntas. Esse modelo ainda permite calcular a distância das palavras do currículo perante os termos pesquisados ou aumentar a relevância de um termo em detrimento a outro. A relevância dos termos pode ser calculada usando os algoritmos BM25 e TF-IDF. Por padrão, a relevância utiliza uma variação do TF-IDF, na qual a relevância é calculada apenas nos currículos que contenham pertinência aos termos digitados.

2.7 BUSCA POR RELEVÂNCIA OU GEOLOCALIZAÇÃO

O protótipo permite ao usuário selecionar dois tipos de busca, por Relevância ou por Geolocalização. Na realidade, se o usuário escolhe a opção Geolocalização, o protótipo realiza busca por relevância e exibe os resultados considerando a posição geográfica de cada profissional. Se o usuário seleciona relevância, os resultados são ordenados usando a relevância dos termos digitados perante o currículo, e considerando também o tipo da busca, que pode ser acadêmica ou profissional.

2.8 BUSCA ACADÊMICA OU PROFISSIONAL

O algoritmo de busca foi criado para permitir a recuperação com enfoque acadêmico ou profissional. Se o usuário seleciona o enfoque acadêmico, o protótipo soma o TF-IDF das colunas Resumo do Currículo, Publicações e Projetos de Pesquisa para calcular a relevância dos currículos perante os termos digitados pelo usuário. Caso usuário selecione o enfoque profissional, o protótipo soma o TD-IDF das colunas Formação, Formação Complementar, Atuação Profissional e Atividades Profissionais para gerar a relevância final do currículo perante os termos digitados.

3 VALIDAÇÃO DO PROTÓTIPO

O protótipo foi apresentado a equipe da PRPPG e foi solicitado que fizessem testes com o sistema de busca do protótipo. O objetivo da avaliação foi validar qualitativamente a utilização do protótipo e sua possível adoção como ferramenta para o setor de programas de pós-graduação da PRPPG. Para evitar exposição, nos resultados, foram apresentados apenas o primeiro nome e

siglas do sobrenome dos servidores. A seguir apresenta-se os resultados dos testes feitos e avaliados juntamente com a equipe da PRPPG:

3.1 BUSCA ACADÊMICA COM O TERMO PECUÁRIA, ORDENADO POR RELEVÂNCIA:

A Tabela 2 exibe os 3 primeiros currículos retornados e a incidência do termo buscado em cada uma das colunas indexadas na busca acadêmica:

Tabela 2 - Resultado da busca do termo pecuária

Servidor	Distância	Resumo	Produções	Proj. Pesquisa
1-Antonio A.R.A.	177.95 km	0	4	1
2-Wanderci A.B.	163.36 km	0	7	0
3-Carlos A.D.S.	68.59 km	0	26	0

Fonte: Elaborado pelo autor.

Notou-se que o servidor exibido na terceira posição aparentemente possui uma maior relevância considerando que já realizou 26 produções no tema pesquisado. O cálculo individual de TF-IDF por coluna pode implicar em possíveis discrepâncias. Foi avaliado o cálculo do TD-IDF no modelo acima e notou-se que para o servidor 1, encontrou-se 5.29 na coluna *Produções* e 3.39 na coluna *Projetos de pesquisa*, já o servidor 3 recebeu 8.17 na coluna *Produções* e 0 na coluna *Projetos de pesquisa*. A servidora 2 recebeu 5.23 na coluna *Produções* e 0 na coluna *Projetos de pesquisa*. Considerando os valores TD-IDF, é possível perceber que o protótipo pode ser alterado para ser mais preciso, pois a relevância é calculada individualmente por coluna, dessa forma, aparentemente o servidor 3 seria o mais relevante no assunto, e foi exibido na terceira posição. O TD-IDF da coluna *Projetos de pesquisa* do servidor 1 pode gerar um valor maior do que o TF-IDF da coluna *Produções* do servidor 3, uma vez que não há normalização entre os dados. Como a relevância é calculada individualmente e o valor de TD-IDF não sofre padronização, que pode prejudicar os resultados. Apesar do método de somatório dos TF-IDF ser utilizado em diversos trabalhos, é possível perceber que calcular a relevância em todas as colunas de uma única vez pode apresentar um resultado mais preciso.

3.2 BUSCA ACADÊMICA COM OS TERMOS PLANTA DANINHA, ORDENADO POR RELEVÂNCIA:

A Tabela 3 exibe os 3 primeiros currículos retornados e a incidência dos termos buscados em cada uma das colunas indexadas na busca acadêmica:

Tabela 3 - Resultado da busca dos termos planta daninha

Servidor	Distância	Resumo	Produções	Proj. Pesquisa
1-Neimar D.F.D	5.76 km	0	28	3
2-Joao P.L.	142.92 km	0	13	0
3-Ana Cardoso C.F.F.D.P.	210.4 km	0	19	0

Fonte: Elaborado pelo autor.

Analisando os resultados, o currículo do servidor 1 é realmente o mais relevante no tema. Como a pesquisa foi realizada com os termos sem aspas, os resultados apresentados consideram tanto o termo **planta** como o termo **daninha**. Uma observação importante é que alguns resultados foram listados, como o servidor **Ricardo S.P.**, que possui em seu currículo a expressão **"planta piloto"**, nesse caso, trata-se de uma polissemia, pois, o protótipo não avalia o contexto para identificar a semântica da palavra. Uma solução poderia ser a utilização de busca semântica baseada em ontologias.

3.3 BUSCA ACADÊMICA COM O TERMO FUZZY, ORDENADO POR GEOLOCALIZAÇÃO:

A Tabela 4 exibe os 3 primeiros currículos retornados e a incidência do termo buscado em cada uma das colunas indexadas na busca acadêmica:

Tabela 4 - Resultado da busca do termo fuzzy

Servidor	Distância	Resumo	Produções	Proj. Pesquisa
1-Amarildo M.D.M.	0.22 km	1	2	0
2-Andre M.K.	69.64 km	0	1	0
3-Luciano S.	156.76 km	0	0	1

Fonte: Elaborado pelo autor.

Foi avaliado o endereço do servidor 1 no sistema ERP do IFMG e realmente é o servidor com endereço mais próximo do local onde foi realizado a busca. A apresentação dos resultados por geolocalização funcionou corretamente, sendo possível ordenar em ordem crescente ou decrescente de distância.

3.4 BUSCA PROFISSIONAL COM O TERMO "REDES DE COMPUTADORES", ORDENADO POR RELEVÂNCIA:

A tabela 5 exibe os 3 primeiros currículos retornados e a incidência do termo buscado em cada uma das colunas indexadas na busca profissional:

Tabela 5 - Resultado da busca dos termos "redes de computadores"

Servidor	Distância	Formação	Atuação	Atividades
1-Cristiano L.C.R.	14.16 km	1	2	1
2-Everthon V.D.S.	184.98 km	2	1	3
3-Chirlando W.D.S.R	170.59 km	3	0	1

Fonte: Elaborado pelo autor.

Os dois primeiros servidores são professores de uma disciplina chamada "Redes de Computadores" na unidade Formiga do IFMG. A busca utilizando os termos entre aspas possibilitou garantir que o protótipo exibisse apenas currículos que realmente possuem alguma experiência nesse tema. O protótipo utilizou corretamente as colunas indexadas na busca profissional para calcular a relevância, entretanto, assim como no exemplo 1, a relevância calculada individualmente pode interferir no cálculo do melhor resultado.

3.5 BUSCAS POR NOMES DE SERVIDORES

Foi realizada uma busca por nome de alguns servidores, como por exemplo, "Luciano S." e o protótipo retornou 6 currículos com esse nome. Cada uma das linhas apresenta quantitativos diferentes nas colunas. Essa inconformidade nos registros ocorreu devido a limitação de unicidade dos currículos. Uma possível solução para o problema de inconsistência de dados, seria o acesso direto à base do Lattes.

3.6 OUTRAS BUSCAS REALIZADAS PELA EQUIPE DA PRPPG

A equipe da PRPPG realizou diversas pesquisas por diferentes temas em diferentes áreas do saber, a seguir apresenta-se o resultado de dois desses testes:

- Busca acadêmica e por relevância pelo termo **"Literatura Brasileira"** => o protótipo retornou currículos da área de literatura, mas também currículos de outras áreas, como enfermagem (ciências biológicas e da saúde), desenvolvimento social, filosofia (área ciências sociais aplicadas).

- Busca acadêmica e por relevância pelo termo **Letras** => A busca por essa palavra retornou produções dessa área, mas também produções da área de matemática, matemática computacional.

Segundo a constatação da equipe, o protótipo retorna de maneira muito precisa os resultados, entretanto, foi apontado por essa equipe como melhoria a possibilidade de realizar filtros por área do conhecimento.

5 CONSIDERAÇÕES FINAIS

É necessário repensar a maneira como as informações internas sobre servidores, processos e instituição são representadas e recuperadas pelos seus colaboradores. É possível melhorar a experiência do usuário com a adoção de técnicas que possibilitem aos mesmos recuperar a informação em uma abordagem mais personalizada, considerando sua geolocalização e hábitos de busca.

Esse trabalho apresentou o desenvolvimento e aplicação de um modelo de RI baseado em busca vetorial e geolocalização para recuperar informações sobre currículos dos servidores do IFMG. A seguir, apresenta-se as principais considerações sobre os resultados, que abrem espaço para trabalhos futuros:

- Como apresentado na seção 3.5.1, o protótipo cria indicadores para cada área do currículo e esse fator é muito relevante para acompanhar tanto a vida acadêmica, quanto profissional do servidor;
- A possibilidade de filtrar por área do conhecimento na busca de currículos poderia melhorar a experiência do usuário;
- O formato de cálculo individual do TF-IDF poderia ser aprimorado caso realizasse o cálculo de todas as colunas indexadas de uma única vez;
- A implementação de técnicas de semântica e ontologias criadas automaticamente ou manualmente poderia melhorar o sistema de RI;
- A inconsistências apresentadas no uso da ferramenta Escavador podem ser resolvidas mediante a celebração de um acordo de **Cooperação Técnica**, conforme o CNPq disponibiliza para instituições de ensino. Esse acordo permite acesso online a base de

dados do Lattes. Para que o IFMG tenha uma precisão e confiabilidade maior nas informações evitando currículos repetidos, é necessário pactuar esse acordo.

- Expandir a base de currículos contemplando os diretórios de grupos de pesquisa disponíveis no Lattes.

O investimento em pesquisa permite um potencial de desenvolvimento contínuo dos programas de pós-graduação, no entanto, é necessária a adoção de um modelo de gestão desses programas fundamentado na utilização de sistemas de informação para tomada de decisão, face ao desafio de proporcionar a sociedade e a academia um retorno efetivo do investimento.

Information Retrieval on Geolocation Curriculum: a prototype using vector search

ABSTRACT

The Federal Minas Gerais Institute (IFMG) received from the federal government R$ 790,575.82 in 2013 and R$ 1,109,576.55 in 2014 to support researchers and research incentives. All this investment requires these programs being audited both the regulatory standpoint, as management. Tools that improve the retrieval of information on researchers and research become important. Considering the lack of such tools in IFMG, the question of this research discusses the improvement of professional and researchers information retrieval processes. Through applied research, we developed a prototype using vector space IR in Lattes curricula IFMG and a ranking was created using geolocation, given that the IFMG employees are located in 17 cities. Finally, the prototype is implemented and evaluated in a qualitative perspective by IFMG research management team. The result shows the importance of geolocation in creating a knowledge organization map.

***Keywords**:* Information retrieval**.** Information retrieval vector space model**.** Lattes**.** Curriculum. IFMG**.** Geolocation**.** Scrapy**.** Crawler**.**

REFERÊNCIAS

BARACHO, M. A. R; CENDÓN, V. B. Esquema de classificação para recuperação de informação em projetos de engenharia. USP, **IX Enancib**, São Paulo, 2008.

BRITO, G. C. A.; QUONIAM, L.; MENA-CHALCO, P. J. Exploração da Plataforma Lattes por assunto: proposta de metodologia, Campinas. **TransInformação**, 77-86, 2016.

CHOO, C. W.; ROCHA, E. A organização do conhecimento: como as organizações usam a informação para criar conhecimento, construir conhecimento e tomar decisões. São Paulo: **SENAC**, 2006.

Escavador. Disponível em: http://www.escavador.com. Acesso em: 20 jun. 2016.

FERRAZ, R. N. R.; QUONIAM, M. L. A utilização da ferramenta computacional ScriptLattes para avaliação das competências em pesquisa no Brasil, **Prisma**, n21, 2013.

GARCIA-MOLINA, H. C. J. The Evolution of the Web and Implications for an Incremental Crawler. Department of Computer Science, Stanford, CA, 1999.

GIL, A. C. **Métodos e técnicas de pesquisa social**. São Paulo: Editora Atlas, 1994. Cap. 3 e 8.

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos de metodologia científica**. São Paulo: Atlas, 3ed, 1991.

MAGALHÃES, J.; QUONIAM, L.; MENA-CHALCO, J.; SANTOS, A. Extração e tratamento de dados na base Lattes para identificação de core competencies em dengue, Londrina. **Informação & Informação**. v. 19, n.3, p. 30-54, 2014.

MARTINS, C. B.; MACCARI, E. A.; STOROPOLI, J. E.; ALMEIDA, M. I. R.; & RICCIO, E. L. A influência do sistema de avaliação nos programas de pós-graduação stricto sensu brasileiro. **Revista Gestão Universitária Na América Latina-GUAL**, 5(3), 155–178.

MORVILLE, Peter; ROSENFELD, Louis. **Information Architecture for the World Wide Web**. Sebastopol, CA: O'Reilly; 1998. ISBN:1-56592-282-4.

MYSQL Documentation. Disponível em: https://dev.mysql.com/doc/. Acesso em 21 jun. 2016.

Portal da Transparência. Controladoria Geral da União, Governo Federal, 2016. Disponível em: http://www.portaldatransparencia.gov.br. Acesso em: 18 jun. 2016.

ScriptLattes. Uma ferramenta para extração e visualização de conhecimento a partir de Currículos Lattes, 2016. Disponível em http://scriptlattes.sourceforge.net/. Acesso em: 20 jun 2016.

SEADLE, M. The new mission of a new i-shcool. **Esmerald Insight.** V.25, iss 1, pp.5-9, 2007

SINGHAL, A. Modern Information Retrieval: A Brief Overview. **Bulletin of IEEE Computer Society Technical Committee on Data Engineering**, 2001.

SOUZA, R. R.; ALMEIDA, B. M.; BARACHO, M. A. R. Ciência da informação em transformação: *Big Data*, nuvens, redes sociais e Web Semântica. **ResearchGate**, Brasília, v. 42, n. 2, p.159-173, 2013.

YADAV, Monika; GOYAL, Neha. Comparrison of Open Source Crawlers – a Review. **International Journal of Scientific & Engineering Research**, v.6, issue 9, 2015.

WERSIG, G. Information science: the study of postmodern knowledge usage, **Information Processing & Management**, v.29, n.2, p.229-239,1993.