CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS TIMÓTEO

Daniel Gomes de Oliveira

BUSCA POR CONTEXTO SOBRE A BÍBLIA SAGRADA UTILIZANDO UM TESAURO GENÉRICO

Timóteo

2018

Daniel Gomes de Oliveira

BUSCA POR CONTEXTO SOBRE A BÍBLIA SAGRADA UTILIZANDO UM TESAURO GENÉRICO

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Douglas Nunes de Oliveira

Timóteo

2018

Daniel Gomes de Oliveira

BUSCA POR CONTEXTO SOBRE A BÍBLIA SAGRADA UTILIZANDO UM TESAURO GENÉRICO

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Trabalho aprovado. Timóteo, 19 de novembro de 2018:

Prof. Me. Douglas Nunes de Oliveira
Orientador

Prof. Me. Deisymar Botega Tavares
Professor Convidado

Prof. Me. Marcelo de Sousa Balbino
Professor Convidado

Timóteo
2017

Agradecimentos

A Deus pela saúde, força e persistência que me tem dado durante o decorrer do curso.

Aos meus pais e namorada Cláudia por todo o apoio nos momentos mais difíceis.

Ao meu orientador Douglas, por todo o suporte e apoio na realização deste trabalho.

E ao restante de familiares e amigos que contribuíram de forma direta ou indireta para minha formação.

Resumo

A Bíblia Sagrada é um livro de extrema importância e um objeto de estudo de religiosos por todo o mundo, portanto, utilizar técnicas de Recuperação de Informação em uma aplicação específica como a bìblia sagrada é um meio importante para se buscar informações que são de interesse dos usuários. Esse trabalho busca verificar a relevância de resultados obtidos em um sistema tesaurico, utilizando sinônimos, antônimos e flexões verbais com o acréscimo de termos radicalizados. Os sinônimos, antônimos e flexões verbais foram recuperados do sistema online Dicio, e a radicalização de termos foi executada pelo algoritmo RSLP (Removedor de Sufixos da Língua Portuguesa). Foi possível perceber que o uso das técnicas mencionadas em uma base bíblica apresentou altos índices de satisfação dos usuários, precisão e revocação médios quando os pesos padrão do sistema foram aplicados aos contextos de buscas. Portanto, concluiu-se que o uso de um tesauro genérico acrescido de termos radicalizados utilizando-se a bíblia como documento base, pode-se retornar resultados relevantes aos usuários.

Palavras-chave: Bíblia Sagrada; Recuperação de Informação; Tesauro.

Abstract

The Holy Bible is a book of paramount importance and an object of religious study throughout the world, therefore, to use Information Retrieval techniques in a specific application such as the sacred Bible is an important means to seek information that is of interest to the users. This work seeks to verify the relevance of results obtained in a thesauric system, using synonyms, antonyms and verbal inflections with the addition of radicalized terms. The synonyms, antonyms and verbal inflections were retrieved from the Dicio online system, and the radicalization of terms was performed by the RSLP (Portuguese Language Suffix Remover) algorithm. It was possible to notice that the use of the mentioned techniques on a biblical basis presented high levels of user satisfaction, average accuracy and recall when the system standard weights were applied to search contexts. Therefore, it was concluded that the use of a generic thesaurus plus radicalized terms using the bible as the base document, can return relevant results to users.

Keywords: Holy Bible; Information Retrieval; Thesaurus.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A área que abrange a Recuperação de Informação (RI) se destina a fornecer aos usuários facilidade de acesso as informações de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2013). Essas informações não necessariamente tendem a ser a busca exata de um determinado usuário, porém, ela deve se manter relevante a pesquisa efetuada. Portanto, um tesauro que engloba os termos e suas propriedades, é uma ferramenta importante para recuperar informações que sejam de interesse do usuário.

A bíblia sagrada é um livro extenso que serve como base de estudos para um grande grupo de pessoas. Portanto, é de extrema importância o uso de ferramentas que auxiliem seus usuários a encontrarem termos similares em um sistema de busca.

Porém, existe uma escassez de aplicações que mantenham uma busca que retornem resultados mediante o contexto de termo-chave. A maioria das aplicações disponíveis gratuitamente mantém pesquisas de termos exatos, não recuperando informações que possam estar relacionadas a busca inserida pelo usuário.

Com a criação de tal ferramenta, o usuário pode não necessitar de utilizar ferramentas de busca genérica, tendo a esse a disponibilidade de uma ferramenta de pesquisa específica sobre a bíblia na língua portuguesa, mediante contexto de termos-chave. A necessidade de busca de uma solução com resultados melhores que busca exata e menos complexa que um buscador justifica este trabalho.

1.1 Justificativa

Este trabalho se justifica da necessidade de criação de uma aplicação que efetue busca por termos relacionados em uma base bíblica na língua portuguesa, que sirva de objeto de estudo para um extenso grupo de estudiosos bíblicos. Para tal, é necessária a utilização de técnicas e métricas de Recuperação de Informação para que dados relevantes sejam apresentadas ao usuário.

Atualmente existe uma gama de aplicações tanto móveis quanto web que servem como ferramentas de pesquisa em uma base bíblica. As buscas efetuadas nesses sistemas apresentam resultados ordenados pela sequência bíblica, exibindo os versículos bíblicos em que tais termos se encontram.

Há também, ferramentas de busca pela web que retornam termos correlacionados aos da busca. Tais ferramentas são capazes de retornar buscas que podem ser de maior relevância para o usuário, devido as técnicas eficazes de indexação e classificação de conteúdo.

Como exemplo podemos citar o trabalho de Bindá, Brandt e Piedade (2013), que utiliza a API Java Lucene como uma forma de indexar os termos bíblicos, fazendo com que os termos exatos a consulta sejam pesquisados e retornados os respectivos versículos bíblicos.

Prover meios de integração de uma ferramenta que utilize um escopo bíblico de informações e de termos correlacionados a pesquisa do usuário, são de extrema importância para os estudiosos bíblicos, devido a alta correlação semântica de termos pela bíblia. Uma forma de unir tais fatos, seria com o uso de um tesauro, que possa indexar termos relacionados sinonimamente, antonimamente, verbalmente e por radicais sintáticos à pesquisa do usuário.

Segundo Baeza-Yates e Ribeiro-Neto (2013), para recuperar respostas de uma consulta, qualquer sistema de RI tem de lidar com um problema central, que é prover quais documentos os usuários irão considerar relevantes e quais irão considerar como irrelevantes. Para tal, é necessário um algoritmo de ranqueamento que se aproxime da opinião do usuário para documentos relevantes. Portanto, uma técnica de ranqueamento eficiente atrelado a um tesauro, podem contribuir para recuperar informações que atendam a necessidade dos usuários.

1.2 Problema

As seguintes questões constituem o problema deste trabalho: A inserção de radical de termos em um tesauro genérico, é uma técnica eficiente na recuperação de termos correlacionados em uma base específica como a bíblia? A utilização de um tesauro genérico em uma base bíblica é um meio eficiente na recuperação de termos relevantes ao usuário?

Um tesauro genérico se constitui de uma base, que apresenta relações de termos variados em uma determinada língua, assim como suas características. O uso do mesmo para constituição de uma base bíblica, serve para verificar a proximidade semântica dos termos pesquisados, com os termos presentes pela base bíblica.

A inserção de radicais em um tesauro consiste no uso de um algoritmo radicalizador, para o agrupamento de termos de mesmo radical. Tal técnica é utilizada para se encontrar termos que sejam etimologicamente próximos a busca do usuário.

A utilização de radicais (*stems*) é muito usada para indexação de termos, fazendo com que se melhore o desempenho em sua recuperação. Baeza-Yates e Ribeiro-Neto (2013) citam que a utilização de *stems* são úteis na melhoria da performance de recuperação, pois reduz as variantes a um termo comum.

A classificação de documentos em Recuperação de Informação, constitui-se na atribuição de pontuação aos documentos da base. O modelo vetorial em específico, considera cada um dos documentos de consulta e de resposta como vetores, e busca a similaridade entre os mesmos. A análise de eficiência na recuperação de termos, consiste na análise qualitativa da recuperação dos termos, ou seja, na comparação dos resultados com a intenção dos usuários.

1.3 Objetivos

Para responder ao problema proposto, o objetivo geral deste trabalho é avaliar se métodos de busca por contexto utilizando tesauro genérico incluindo termos radicalizados, retornam buscas relevantes aos usuários. A relevância dos documentos buscados se dá quando os re-

sultados estão próximos ao que o usuário quer como resposta. A proximidade do que quer o usuário com os resultados obtidos, se dá através de métricas de análise qualitativa, descritos com maior detalhes na seção 5.2.

Também, objetivam-se mais especificamente:

1. Desenvolver uma ferramenta de busca sobre a base bíblica local;
2. Adaptar o modelo de recuperação vetorial para a busca de termos correlacionados em uma base por tesauro;

1.4 Estrutura da monografia

Este trabalho foi organizado em seis capítulos. Os capítulos são apresentados em uma ordem sequencial de eventos, com vista a facilitar o entendimento sobre sua elaboração.

O capítulo 2 apresenta os passos metodológicos de elaboração deste trabalho. Tais passos descrevem o processo de seleção de ferramentas de coleta das informações utilizadas no tesauro, a de escolha de um banco de dados, e o planejamento de avaliação do sistema.

O capítulo 3 apresenta a bibliografia que contém as técnicas as quais se propõe se utilizar neste trabalho. Tais técnicas estão relacionadas ao uso de tesauros, a utilização de radicalizadores e sobre a modelagem de um sistema de RI mediante ao ranqueamento de conteúdo.

O capítulo 4 apresenta a forma de construção da ferramenta que servirá de avaliação. Todas as técnicas referentes a construção do banco de dados com os dados indexados e da ferramenta de uso, serão aprofundadas com mais detalhes neste capítulo.

O capítulo 5 apresentará o detalhamento dos resultados referentes a construção da ferramenta, e dos dados de avaliação por parte dos usuários.

O capítulo 6 apresentará a conclusão deste trabalho, definindo se o uso de tais técnicas podem retornar resultados relevantes a pesquisa. Também serão apresentadas algumas sugestões de trabalhos futuros.

2 Procedimentos metodológicos

Neste capítulo são apresentados os métodos e instrumentos utilizados para a realização dessa pesquisa.

Segundo, Gerhardt e Silveira (2009) a pesquisa qualitativa se preocupa com o aprofundamento da compreensão de um grupo social, de uma organização, etc. Portanto, essa pesquisa é qualitativa devido a análise de relevância dos documentos buscados pelos usuários quando comparados a intenção dos mesmos. De acordo com Fonseca (2002), a pesquisa quantitativa recorre à linguagem matemática para descrever as causas de um fenômeno, as relações entre variáveis, etc. Portanto, essa pesquisa também é quantitativa devido a análise dos índices de avaliação gerados pelos resultados das pesquisas.

Segundo Gerhardt e Silveira (2009), a pesquisa exploratória tem como objetivo proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses. Portanto, quanto aos objetivos essa pesquisa é exploratória, devido ao levantamento de informações sobre tesauros, radicalizadores e métodos de avaliação de Recuperação de Informação, de forma a integrá-los e utilizá-los na ferramenta de avaliação dos usuários.

Quanto aos procedimentos, essa pesquisa é um estudo de caso, por fazer uma análise de um tesauro genérico em uma base específica, sendo essa a bíblia sagrada. Segundo Gerhardt e Silveira (2009), um estudo de caso pode ser caracterizado como um estudo sobre uma entidade bem definida.

Os procedimentos metodológicos são organizados nas etapas abaixo:

1. reunir e estudar os trabalhos que envolvem análise de técnicas de recuperação de informação em uma base de dados, para servir como meio de auxílio para os procedimentos e técnicas a serem adotados na presente pesquisa.

2. levantar sistemas de tesauro genérico disponíveis gratuitamente pela web, para fornecer os termos sinônimos, antônimos, e de conjugações verbais sobre os termos bíblicos, e radicalizar tais termos com o auxílio de um algoritmo radicalizador.

3. projetar e implementar um banco de dados que reúna todos os termos bíblicos, assim como os termos antônimos, sinônimos e de conjugações verbais, de forma a indexá-los para maior completude da pesquisa do usuário;

4. implementar um software para servir de ferramenta de buscas e de avaliação por parte do usuário;

5. avaliar os resultados propostos aos usuários, de forma a apresentar dados de relevância das buscas efetuadas, e gerar uma avaliação crítica sobre o sistema com base em tal avaliação.

As etapas citadas serão detalhadas ao decorrer do capítulo para maior compreensão das mesmas.

A revisão de literatura realizada para este trabalho, consiste no levantamento bibliográfico das técnicas a que se propõe utilizar neste trabalho. Portanto, as técnicas de utilização de tesauros, radicalizador e meios de ranqueamento de documentos, foram analisados separadamente de forma a reuni-las e identificar quais delas serão úteis na realização deste trabalho.

O levantamento das técnicas relacionadas ao trabalho, é apresentado no capítulo 3, com conceitos e abordagens específicas presentes na literatura. Foram selecionadas abordagens tanto na área de Ciência da Computação aplicada a área de RI, quanto na área de Biblioteconomia. Tais abordagens foram analisadas, e determinadas quais destas técnicas podem ser eficazes neste trabalho. Tais técnicas foram selecionadas mediante a hipótese de eficácia referentes aos seus conceitos abordados e que podem servir como meio para a eficiência de resultados para este trabalho.

2.1 Utilização de tesauro genérico

Para a construção do trabalho, é necessário a utilização de um tesauro para a inserção de termos correlacionados. Buscou-se por tesauros na língua portuguesa disponibilizados pela web, que atendesse com completude os elementos necessários para este trabalho. Dentre esses elementos estavam tesauros que tivessem a maior quantidade de termos presentes na bíblia, e que tivesse em sua base a relação de sinônimos, antônimos e de conjugações verbais.

Dentre os tesauros encontrados, estão: o Sinônimos-Online (ZIRKELBACH, 2012-2016), o Woxikon (EISBäR MEDIA GMBH, 2016) e o Dicio (7GRAUS, 2009-2016).

Após analise dos tesauros apresentados, verificou-se que o tesauro Sinônimos-Online apresentava apenas os sinônimos e conjugações verbais, não contendo a análise de antônimos. Para utilizar a conjugação verbal da base Sinônimos-Online e do Woxicon, deve-se saber que o termo inserido é verbo, portanto, para a utilização sobre os termos bíblicos seria necessário saber a classe gramatical dos mesmos, o que não se propõe este trabalho.

Mediante os fatos apresentados, a base Dicio foi a que mais se adequou a necessidade deste trabalho. A base Dicio está disponível na web e gratuitamente. A mesma já contém uma API não oficial de acesso aos dados implementada, a *Unofficial Python API for Dicio*, que foi proposta por Felipe Pontes e está disponível em *https://github.com/felipemfp/dicio* (PONTES, 2016).

Um fator a ser testado neste trabalho é se a forma raiz dos termos utilizando um radicalizador, é uma estratégia eficiente como integração semântica em um tesauro. Portanto, deve-se radicalizar os termos presentes na base de dados. Neste trabalho será utilizado o algoritmo radicalizador RSLP, proposto por Orengo e Huyck (2001) e aprimorado por Coelho (2007), que é um algoritmo feito para a língua portuguesa. Para Xavier, Gomes e Silva (2013), as vantagens de se utilizar o RSLP é que além de analisar diversas regras específicas da língua portuguesa, ele ainda conta com um dicionário de exceções.

Os conceitos sobre tesauro e radicalizadores, se encontram no capítulo 3.

2.2 Projeto de Banco de Dados

Este trabalho propõe a utilização de um tesauro bíblico, com dados coletados de um tesauro genérico, acrescido da técnica de radicalização dos termos. Todos estes dados coletados, devem estar em um banco que mantenha a correlação dos mesmos, assim como os dados de acesso de um usuário específico, que servirá como base para os dados de avaliação da aplicação.

Para este trabalho optou-se pela utilização do banco de dados relacional MySQL devido a sua licença de software livre e a sua alta velocidade na execução de consultas. Segundo Milani (2007), o MySQL é altamente veloz pelo fato da linguagem ter sido implementada por meio de códigos e funções otimizadas pelos seus desenvolvedores.

2.3 Software de Busca

O software de busca pelos termos bíblicos, é que permite a interação dos avaliadores com o sistema. Portanto, é essencial que o software seja prático e intuitivo para gerir informações úteis sobre buscas relevantes por parte de tais usuários. A utilização de uma aplicação web, faz com que o sistema se mantenha com maior disponibilidade e inter-operabilidade devido ao uso em diversos Sistemas Operacionais.

Para este trabalho, foi desenvolvido uma ferramenta web utilizando-se a linguagem Ruby junto com o framework Ruby on Rails. Tal ferramenta foi integrada junto ao radicalizador de termos implementado em linguagem C, através de um arquivo executável a ser chamado pelo projeto principal.

No sistema proposto, é de vital importância a disposição dos resultados de busca para os usuários, de forma a apresentar os documentos mais relevantes no topo da lista de resultados. Portanto, no software será implementado o algoritmo de ranqueamento baseado no modelo vetorial, que constitui-se de um dos modelos clássicos na área de RI. No entanto, o sistema será adaptado de forma a indexar os termos pelo tesauro proposto. Os conceitos referentes ao ranqueamento presentes na literatura sobre modelos de RI, estão descritas no capítulo 3.

As principais vantagens de utilização do modelo vetorial descritas por Baeza-Yates e Ribeiro-Neto (2013), são:

1. Seu esquema de ponderação de termos que melhora a qualidade de recuperação;
2. Pode retornar resultados parciais às buscas realizadas;
3. A utilização da fórmula do cosseno, que atribui o grau de similaridade do documento a busca;
4. A normalização pelo tamanho do documento está embutida na fórmula;

2.4 Avaliação do Sistema proposto

Para Baeza-Yates e Ribeiro-Neto (2013), a avaliação de um Sistema de RI advém da necessidade de medir o quão bem o sistema atende a necessidade de informação do usuário, no entanto, cada usuário possui uma visão diferente em relação as buscas retornadas de uma pesquisa. Baeza-Yates e Ribeiro-Neto (2013) citam que é possível definir métricas aproximadas que, na média, tenha correlação com as preferências dos usuários.

As métricas adotadas neste trabalho se baseiam nos métodos de precisão e revocação, que estão detalhados na seção 5.2. A decisão de utilização de tais métodos se basearam na vantagem de utilização dos mesmos citadas por Baeza-Yates e Ribeiro-Neto (2013), que diz respeito a simplicidade dos cálculos realizados pelos métodos.

A avaliação será gerada mediante o uso do sistema por usuários que tenham algum conhecimento bíblico. As buscas tenderão a abrangir todo o arcabouço de termos bíblicos, colocando em teste as técnicas e o meio de ranqueamento apresentados. A relevância dos resultados apresentados estão diretamente relacionados a satisfação dos usuários com a busca, portanto, os usuários responderão a um questionário ao final de cada busca para saber se os resultados atenderam a sua intenção de busca.

3 Revisão da literatura

Neste capítulo são apresentados os conceitos teóricos utilizados, assim como os trabalhos que estejam relacionados a esse. Foram explicitadas as técnicas predominantes para a relação de termos pela base existente, assim como para sua recuperação através de uma busca. Tais técnicas foram apresentadas por revisão literária de trabalhos de outros autores, como forma de compreensão de tais técnicas neste trabalho.

Este capítulo se encontra dividido em quatro seções. A primeira seção aborda o Estado da Arte do trabalho proposto, abordando alguns dos trabalhos que estão relacionados a este. A segunda seção faz uma abordagem técnica sobre conceitos de tesauros, e seu uso para recuperação de termos correlacionados semanticamente. A terceira seção diz a respeito a radicalização de termos, e sobre o uso do algoritmo RSLP. E a quarta seção fala sobre a técnica de ranqueamento de busca a que este trabalho se propõe a utilizar.

3.1 Estado da Arte

Este trabalho consiste na avaliação de uso de um tesauro de uso geral aplicado a uma base específica como a bíblia. Vários conceitos para o aprimoramento da base tesaurica e da forma de ranquear tal conteúdo foram apresentados. Segue-se a definição de alguns dos trabalhos que utilizam das principais técnicas propostas neste trabalho.

O uso de tesauros como forma de expansão de consultas, tem sido abordado em alguns trabalhos como ferramenta importante na recuperação de documentos que possam estar relacionados a busca do usuário. No trabalho de Antunes (2010), é proposto um Sistema de Recuperação de Informação na área jurídica, com a utilização de um tesauro específico jurídico como forma de expandir as consultas formuladas pelos usuários. Para a criação do mesmo, foi gerada uma ferramenta baseada na ferramenta LUCENE com a utilização do modelo vetorial em RI, e com a utilização do Tesauro Jurídico do Conselho de Justiça Federal (CJF) para expansão automática de consulta.

Este trabalho difere do trabalho proposto por Antunes (2010), pois o mesmo se propõe a utilização de um tesauro de uso geral em uma base bíblica, devido a heterogeneidade que há pelos livros bíblicos. Outra diferença fundamental está na indexação dos termos aos documentos, onde no trabalho de Antunes (2010) é proposto a indexação automática dos termos utilizando a ferramenta LUCENE, e neste trabalho é proposto a indexação em um banco de dados relacional, correlacionando os termos próximos semanticamente aos documentos.

No trabalho de Greenberg (2001), é apresentada uma avaliação sobre sistemas de expansão de consultas baseada em um tesauro de sinônimos sobre o vocabulário controlado ProQuest. Em tal trabalho, utilizou-se para a indexação termos sinônimos parciais, genéricos, específicos e relacionados como forma de se saber se a utilização dos mesmos retornaria resultados mais eficazes na recuperação de documentos relevantes. A forma de pontuação

de ranqueamento foi gerada pelo Coeficiente de Dice, que atribui pontos a similaridade entre duas amostras. No geral, o trabalho mostrou que houve um aumento no parâmetro *recall* e decaimento no parâmetro *precision* das buscas.

Neste trabalho, é proposto um sistema de indexação de termos por sinônimos, antônimos, radicais e conjugações verbais, diferindo assim do trabalho proposto por Greenberg (2001). Outra diferença está no ranqueamento de conteúdo, que será feito utilizando o modelo clássico vetorial em RI, com adaptações ao uso de termos correlacionados.

Outro trabalho que pode ser citado é o de Voorhees (1993), onde se executou a expansão do vetor de consultas sobre a base TREC-1, utilizando termos sinônimos da base da WordNet. A seleção de tais termos sinônimos foram feitos em um processo particular de escolhas do autor, aumentando assim a eficácia de recuperação se comparados a um processo automático.

Este trabalho difere do trabalho de Voorhees (1993), quanto a utilização da base de coleção já que neste trabalho é proposto um sistema sobre uma base bíblica, e também difere na indexação de termos correlacionados, pois os mesmos são recuperados de uma base genérica sem o processo de seleção particular, e adicionados a um tesauro de termos correlacionados.

No trabalho de Bindá, Brandt e Piedade (2013) é proposto um sistema com aplicação do algoritmo de ranqueamento em uma plataforma Android, utilizando a API Java Lucene como ferramenta de busca na base de dados contendo os textos bíblicos. Para recuperar os documentos, utilizou-se da função de similaridade entre documentos baseado nos modelos vetorial e booleano.

Este trabalho se difere do apresentado por Bindá, Brandt e Piedade (2013) quanto a recuperação dos textos bíblicos, pois neste trabalho é utilizado uma base de termos correlacionados por tesauro, como forma de aumentar a eficácia de recuperação de documentos. Outra diferença está no modelo de ranqueamento dos documentos, onde neste trabalho é proposto a utilização do modelo vetorial, enquanto no de Bindá, Brandt e Piedade (2013) utiliza-se da adaptação entre o modelo vetorial e booleano.

Até onde se observou, não houveram trabalhos que utilizassem de um algoritmo de radicalização de termos para a indexação de forma semântica em um tesauro. Portanto, este trabalho se difere dos apresentados pela literatura quanto a utilização do mesmo em uma base de termos correlacionados, como forma de recuperação de documentos relevantes.

3.2 Tesauro

O termo Tesauro tem origem no latim pela palavra *thesaurus* e pela palavra grega *thesaurós* e significa tesouro (KNAPP, 2000). Portanto, um tesauro em um conceito mais abrangente, pode ser definido como um tesouro ou depósito de palavras. Porém, esse tesouro não se resume apenas a um grande arcabouço de termos independentes:

> "Por tesouro de palavras não se entende apenas quantidade de palavras, mas

> sim riqueza de conceitos e relações semânticas que devem existir entre elas." (GRANADA, 2011, p. 21)

Um tesauro é uma base composta por palavras e sua base significante. Baeza-Yates e Ribeiro-Neto (2013) dizem que um tesauro em sua forma mais simples, consiste de uma lista pré-compilada de palavras importantes em um dado domínio, e para cada palavra dessa lista, um conjunto de palavras relacionadas. A forma mais simples de utilização de tesauro é com a utilização de termos chaves com sua base sinônima de utilização.

Tesauro é amplamente utilizado na área de biblioteconomia como forma de organização de conteúdo literário pelos bibliotecários. Porém, Tal técnica é amplamente empregada na área de RI, como forma auxiliar de recuperação de dados indexados em uma base correlacionada. Campos e Gomes (2006) dizem que um tesauro é uma linguagem documentária, que possui termos correlacionados, com objetivo de indexação/recuperação em um sistema de recuperação de informação. Para Granada (2011), assim como o tesauro ajuda o bibliotecário a encontrar documentos por palavras-chave, ele também auxilia o usuário da internet a encontrar os documentos dos quais necessita.

A popularização de utilização de um tesauro, se deu pelo trabalho de Peter Mark Roget em *Thesaurus of English Words and Phrases* (ROGET, 1911). Em seu trabalho, foi criado um tesauro manual, onde as palavras não eram agrupadas por ordem alfabética como nos demais dicionários presentes na época, mas os termos eram agrupados por ideias expressas. Baeza-Yates e Ribeiro-Neto (2013) citam que Roget expressa as ideias de uma maneira mais complexa do que uma simples representação de ideias por termos sinônimos, pois ele apresenta frases também como conteúdo correlacionado.

Baeza-Yates e Ribeiro-Neto (2013) apresentam um exemplo de entrada de determinado termo do tesauro de Roget:

Figura 1 – Entrada de termo pelo tesauro de Roget

cowardly *adjective*
Ignobly lacking in courage: *cowardly turncoats.*
Syns: chiken (slang), chiken-hearted, craven, dastardly, faint-hearted, gutless, lily-livered, pusillanimous, unmanly, yellow (slang) yellow-bellied (slang).

Fonte: (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 216)

A representação de termos em uma base por tesauro se dá por um conjunto de características pertencentes ao termo apresentado, tanto de forma geral, quanto de forma específica. Tesauros de conteúdos específicos, possuem termos de determinado domínio do saber. Um exemplo citado por Baeza-Yates e Ribeiro-Neto (2013) é do *Thesauros of Engineering and Scientific Terms*, sendo um tesauro de cunho específico a termos da Engenharia. Outro exemplo a ser citado é do *Tesauro eletrônico do cinema brasileiro* (MOURA et al., 2008), que possui termos relacionados ao mundo cinematográfico.

Outro tesauro manual bastante referenciado em literaturas de RI é o da WordNet[1], onde classes gramaticais são agrupados por sinônimos. Segundo Granada (2011), os sinônimos são interconectados através de relações semântico-conceituais e lexicais, e tais relações variam de acordo com o tipo de palavra.

A construção manual de um tesauro por ser um procedimento pioneiro na área, necessita de um profissional especialista para definição de termos. Tal procedimento se torna trabalhoso e exige extrema dedicação. Mediante a isso, surge-se a construção automática de tesauros que se baseia na identificação automatizada de relacionamentos semânticos entre as palavras, encontrando palavras mais similares a uma palavra-chave (GRANADA, 2011).

Como exemplo de construção de tesauro automático podemos citar o trabalho de Ponzetto e Strube (2007), que se baseia na coleta e definição de termos por uma rede semântica de relações pelo Wikipédia.

3.3 Radicalização

Por vezes, o usuário quando executa a busca em um documento com palavras derivadas de uma forma comum, pode não ter resultado algum retornado devido a pequenas diferenças sintáticas existentes entre tais formas. Esse problema pode ser resolvido, colocando todos os elementos variantes em um termo comum, ou seja, em uma mesma raiz.

A radicalização de termos pode ser útil em situações onde existam palavras derivadas por plurais, conjugações verbais, gerúndios, e que podem ser colocadas em uma mesma forma. Um exemplo a ser citado é da palavra “amor”, que possui variantes como “amar”, “amarás”, “amando”, entre outros. Tais termos podem ser colocados em uma forma comum, eliminando todos os seus afixos e colocando-os sob uma forma comum, sendo a raiz ou *stem* “am-”.

Essa diminuição de termos a um termo comum, traz significantes melhorias de desempenho na indexação de tais termos em uma base. Baeza-Yates e Ribeiro-Neto (2013) citam que a utilização de *stems* são úteis na melhoria da performance de recuperação, pois reduz as variantes a um termo comum, e reduz o tamanho da estrutura de indexação, pois diminue a quantidade de termos distintos.

O uso de radicalizador, além de melhorar o desempenho, também pode ser útil na recuperação de termos relevantes a pesquisa do usuário. De acordo com Coelho (2007), ao se reduzir os termos em seus respectivos radicais melhora-se as chances de recuperar um documento desejado pelo usuário, ou seja, melhora-se sua revocação.

Porém, a utilização de radicalizador também traz algumas desvantagens, que são citadas por Coelho (2007) sobre a perca de precisão na recuperação da informação, uma vez que se perde o termo exato por sua raiz e também na indexação de termos homônimos em um mesmo radical. Como exemplo de termos homônimos radicalizados temos o termo “cedo”, que pode ser tanto conjugação do verbo “ceder”, quanto o advérbio de tempo.

[1] http://wordnetweb.princeton.edu/perl/webwn

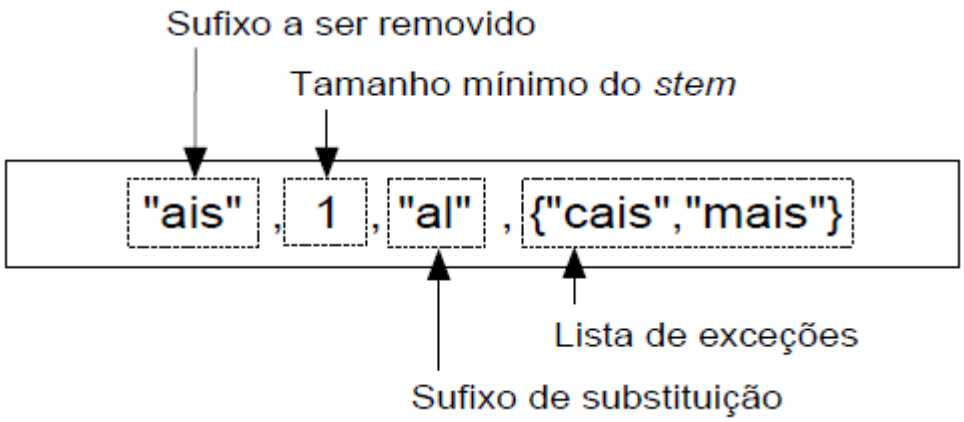
Grande parte dos algoritmos radicalizadores implementados foram feitos para a língua inglesa, e boa parte dos mesmos foram adaptados para a língua portuguesa, um exemplo é o algoritmo de Porter (PORTER, 1980). Devido a isto, tais algoritmos não possuem boa eficiência na radicalização de termos quanto os algoritmos nativos da língua portuguesa. Coelho (2007) diz que a língua portuguesa possui inflexões que causam grandes alterações nos radicais das palavras para impedir a conflação dos termos, e que devido a isso, a utilização de radicalizador nativo a língua é essencial.

3.3.1 Removedor de Sufixos da Língua Portuguesa (RSLP)

Este algoritmo de radicalização de termos foi proposto por Orengo e Huyck (2001), e aprimorado por Coelho (2007), e é um algoritmo nativo a língua portuguesa. O RSLP é um algoritmo similar ao Porter, onde uma série de regras de remoção de sufixos são impostas sucessivamente até se chegar a um termo raiz. Porém, o RSLP possui uma grande quantidade de regras se comparado ao Porter, e ainda conta com uma lista de exceções especialmente feita para a língua portuguesa. Coelho (2007) em seu trabalho efetuou melhorias sobre o algoritmo, adicionando novas funcionalidades como a criação de um dicionário de nomes próprios.

O algoritmo possui em cada passo de remoção um conjunto de regras a serem seguidas, como pode ser visto pela figura 2.

Figura 2 – Exemplo de regra do RSLP



Fonte: (COELHO, 2007, p. 19)

O algoritmo conta com 8 passos de redução executados sequencialmente. Coelho (2007) os define da seguinte forma:

1. **Redução do Plural:** Retira-se o sufixo relacionado ao plural das palavras, normalmente sendo “s”. Nem todos os termos terminados com “s” são plurais, portanto, tais termos devem estar contidos na lista de exceções. Este passo contém 11 regras definidas.

2. **Redução do Feminino:** Neste passo as palavras de gênero feminino são passadas para o masculino. Para este passo, a palavra deve terminar com a letra “a”. Palavras como “casada” serão passadas para “casado”. Este passo contém 15 regras definidas.

3. **Redução Adverbial:** Este passo reduz apenas os advérbios terminados em “mente”, por ser o único que denota advérbio em português. Portanto, palavras como “gravemente” serão reduzidas a “grave”. Este passo possui apenas 1 regra.

4. **Redução do Aumentativo:** Este passo reduz as palavras do aumentativo, superlativo e diminutivo para a forma normal. Portanto, termos que terminam com “-inho” por exemplo, ficarão a uma forma base de raiz. Este passo possui 23 regras definidas.

5. **Redução Nominal:** Este passo foi definido por Orengo e Huyck (2001) com 61 sufixos de adjetivos e substantivos. Porém, quando sua implementação foi feita, ficaram definidos de forma final 84 sufixos. Sufixos comuns como “-ável” e “-ência” foram definidos. Portanto, palavras como “aparência” são reduzidas ao termo “apar”. Com a execução desse passo, os passos 6 e 7 não são executados.

6. **Redução Verbal:** Neste passo são definidas 101 regras definidas, assumindo-se a variedade de formas verbais presentes na língua portuguesa. Verbos como “amarei” são reduzidos a forma “am” nesse passo. Com a execução desse passo, o passo 7 não será executado.

7. **Remoção de Vogais:** Este passo tem o objetivo de remover as vogais (“a”, “e” ou “o”) dos termos que não foram modificados pelos passos 5 e 6. A palavra “casa” não teria seu sufixo removido nos 2 passos anteriores, portanto, nesse passo ela teria a letra “a” retirada e ficaria como “cas”, tendo a mesma forma da redução do termo “casinha”.

8. **Remoção de Acentos:** Neste passo são retirados todos os acentos da palavra, para garantir que a raiz resultante se mantenha da mesma forma que suas variantes. Assim, termos como “geográfica” e “geografia” são reduzidos a uma raiz comum, sendo “geograf”.

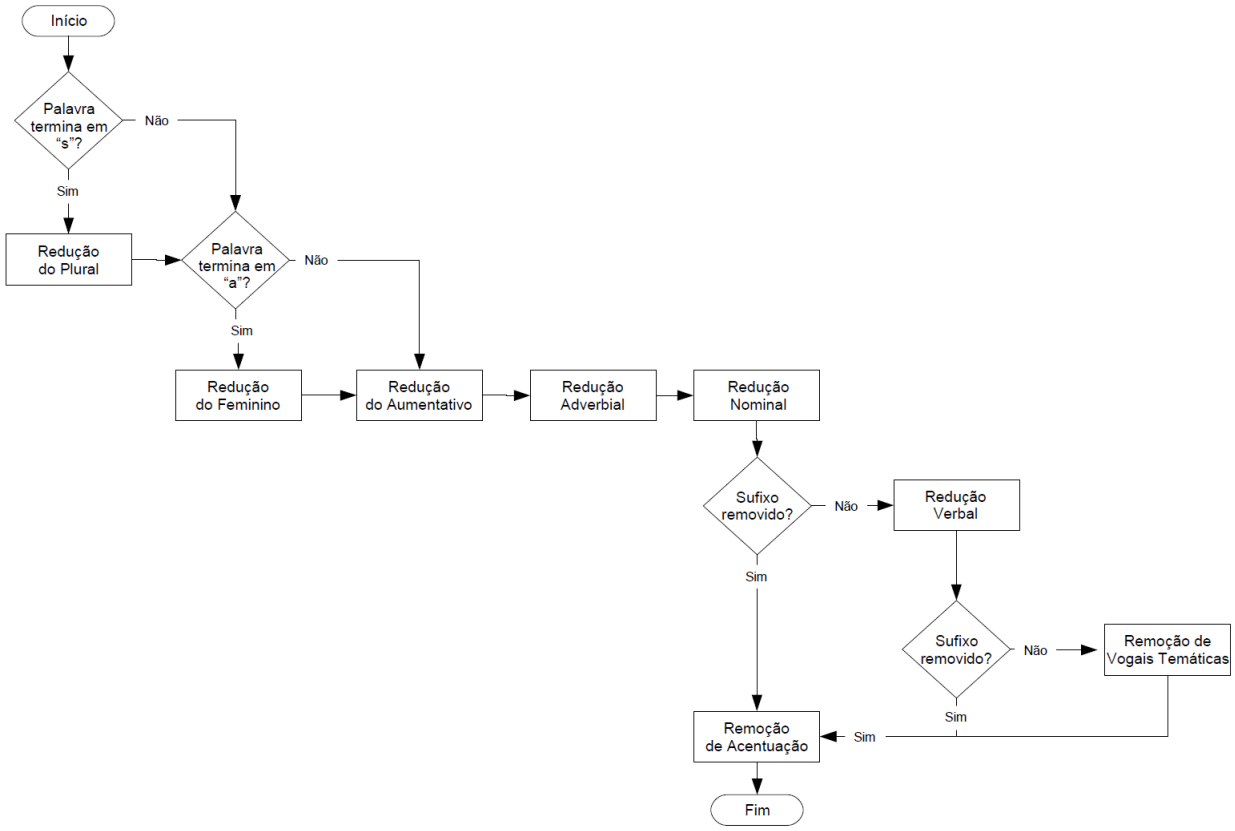
Na figura 3, é apresentado um fluxograma contendo os passos de execução sequencial do algoritmo RSLP.

3.4 Ranqueamento

O sistema de ranqueamento de um sistema em RI é composto por um modelo de recuperação baseado em escores, ou seja, é atribuído a cada documento uma pontuação. Portanto, um modelo de RI define uma função de ranqueamento, colocando em ordem os documentos de acordo com seu grau de relevância.

Baeza-Yates e Ribeiro-Neto (2013) citam em seu livro que os algoritmos de ranqueamento estão no cerne dos sistemas de RI, devido ao fato dos mesmos estarem diretamente relacionados a opinião dos usuários quanto aos documentos considerados relevantes. Baeza-Yates e Ribeiro-Neto (2013) dizem ainda que tal relevância é relativa, já que os usuários possuem opiniões diferentes sobre resultados a uma mesma consulta, porém, um algoritmo de ranqueamento eficiente é o que classifica os documentos de acordo com grande parte dos usuários.

Figura 3 – Sequência de passos do RSLP



Fonte: (ORENGO; HUYCK, 2001, p. 187)

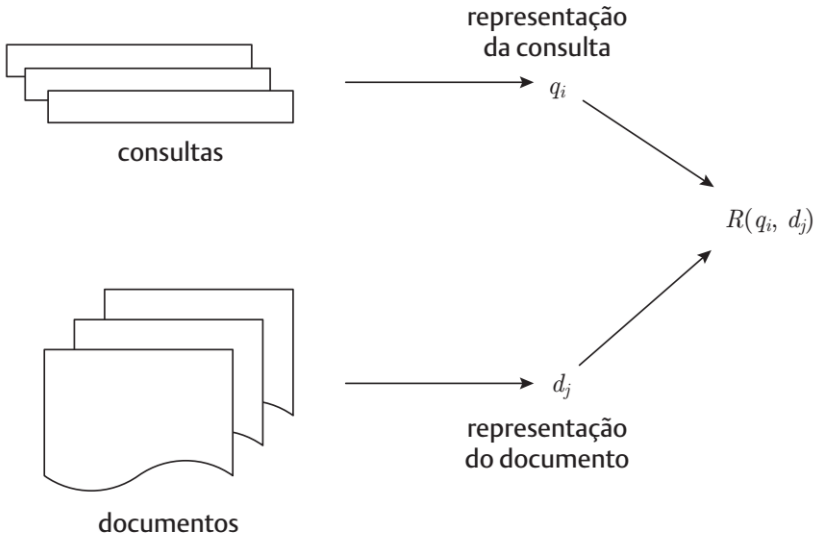
De acordo com Baeza-Yates e Ribeiro-Neto (2013), um modelo de recuperação é uma quádrupla [D, Q, F, $R(q_i, d_j)$] onde

1. D é um conjunto composto por visões lógicas dos documentos da coleção.
2. Q é um conjunto composto por visões lógicas das necessidades de informação dos usuários.
3. F é um arcabouço para a modelagem das representações de documentos, consultas e de seus relacionamentos.
4. $R(q_i, d_j)$ é uma função de ranqueamento que associa um número de uma consulta q_i pertencente a Q, à um documento d_j pertencente a D. Esta função é que define a ordem dos documentos em relação a consulta q_i.

A figura 4 mostra uma representação esquemática das consultas e documentos em um modelo de RI, com a função de ranqueamento representando um elo entre os mesmos, que atribui um grau de similaridade entre uma consulta e um documento através de pontuação.

Existem uma série de modelos de RI formados durante os anos e que são parte importante na área de Recuperação de Informação. Segundo Ferneda (2003), a maioria deles são

Figura 4 – Representação de um modelo de RI



Fonte: (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 23)

de natureza quantitativa, baseada em disciplinas de lógica, estatística e de teoria de conjuntos. Neste trabalho será destacado o modelo vetorial, que é um dos modelos clássicos de RI.

3.4.1 Modelo Vetorial

O modelo vetorial se propõe a atribuir pesos aos termos de indexação da consulta e dos documentos, recuperando documentos que não necessariamente são exatos a busca do usuário. De acordo com Ferneda (2003), os pesos serão utilizados para o cálculo de similaridade da função de ranqueamento, atribuindo a pontuação correspondente de cada documento aos termos de busca inseridos pelo usuário.

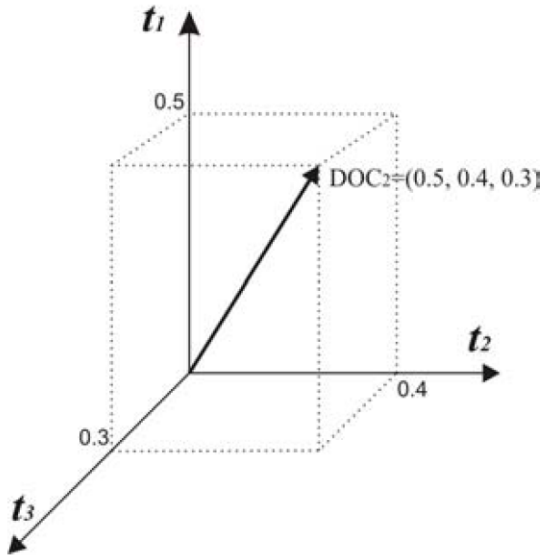
Ferneda (2003) cita que no modelo vetorial, cada documento é representado como um vetor que possui termos de indexação com pesos, que determina o grau de relevância de cada termo em um documento. Tais pesos são normalizados com valores entre 0 e 1, de forma que os pesos com valores mais próximos de 1 indicam ser os termos mais relevantes em tal documento. A figura 5 mostra a representação do vetor de pesos de um documento em um espaço tridimensional.

Ferneda (2003) cita ainda que no modelo vetorial, uma expressão de busca do usuário também é vista como um vetor em um espaço multidimensional. Portanto, são atribuídos pesos também aos termos de indexação da consulta. A figura 6 mostra uma expressão de busca juntamente com dois documentos e seus pesos dos termos de indexação.

Portanto, o modelo vetorial calcula o grau de similaridade existente entre o vetor de expressão de busca do usuário com o vetor correspondente ao documento, atribuindo assim um valor numérico aos mesmos, que servirá de base para o ranqueamento de documentos ao usuário. Baeza-Yates e Ribeiro-Neto (2013) citam que essa relação é quantificada através do

Figura 5 – Representação vetorial de um documento com três termos de indexação

	t_1	t_2	t_3
DOC_2	0.5	0.4	0.3

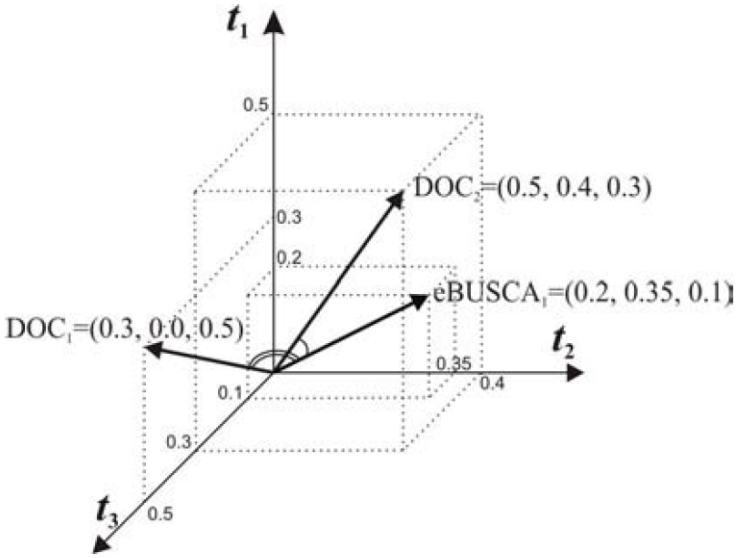


Fonte: (FERNEDA, 2003, p. 28)

Figura 6 – Representação de uma expressão de busca em um espaço vetorial

	t_1	t_2	t_3
$eBUSCA_1$	0.2	0.35	0.1

	t_1	t_2	t_3
DOC_1	0.3	0.0	0.5
DOC_2	0.5	0.4	0.3



Fonte: (FERNEDA, 2003, p. 29)

cosseno do ângulo entre o vetor de consulta q e o vetor do documento d_j. O valor desse grau de similaridade é definido por:

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

Segundo Baeza-Yates e Ribeiro-Neto (2013), a definição dos pesos referentes aos termos dos vetores é dada pela combinação da interpretação estatística da especificidade dos termos juntamente com a frequência dos mesmos. Então,

$$w_{i,q} = (1 + log\ f_{i,q}) \times log\left(\frac{N}{n_i}\right)$$
$$w_{i,j} = (1 + log\ f_{i,j}) \times log\left(\frac{N}{n_i}\right)$$

sendo $f_{i,q}$ a frequência com que o termo aparece na expressão de consulta q, $f_{i,j}$ a frequência do termo no documento d_j, N o número de documentos da coleção e n_i o número de documentos em que o termo k_i aparece.

Utilizando o exemplo da figura 6 definido por Ferneda (2003), o cálculo de similaridade se dará por:

$$sim(DOC1, eBUSCA1) = \frac{(0.3\times0.2)+(0.0\times0.35)+(0.5\times0.1)}{\sqrt{(0.3)^2+(0.0)^2+(0.5)^2}\times\sqrt{(0.2)^2+(0.35)^2+(0.1)^2}} = 0.45$$
$$sim(DOC2, eBUSCA1) = \frac{(0.5\times0.2)+(0.4\times0.35)+(0.3\times0.1)}{\sqrt{(0.5)^2+(0.4)^2+(0.3)^2}\times\sqrt{(0.2)^2+(0.35)^2+(0.1)^2}} = 0.92$$

Portanto, o $DOC2$ possui 92% de similaridade com o $eBUSCA1$ enquanto o $DOC1$ possui 45%. Logo, em um sistema o $DOC2$ seria ranqueado no topo da lista de documentos.

4 Desenvolvimento

Este capítulo trata de explicitar a construção da ferramenta que servirá de base de avaliação pelos usuários, assim como as técnicas e ferramentas de desenvolvimento e análise de dados utilizadas neste trabalho. Serão detalhadas as etapas de construção do sistema, de forma cronológica e construtiva, para melhor compreensão do sistema.

4.1 Construção do tesauro genérico

O sistema é composto por um tesauro genérico que contém informações de sinônimos, antônimos e flexão verbal sobre os termos bíblicos. Para a recuperação de tais informações, foi-se necessário o levantamento dos dados provenientes de um sistema que já obtivesse tais informações.

O sistema em questão é o Dicio (7GRAUS, 2009-2016), que é um sistema online que retornam os resultados mediante pesquisa pelos usuários. De acordo com a definição do sistema por (7GRAUS, 2009-2016), este é um sistema de dicionário de português contemporâneo para uso e estudo da língua portuguesa, portanto é um sistema que mantém dados relevantes para o uso de termos genéricos.

Para recuperação dos termos inseridos em tal sistema, utilizou-se de um programa de busca na linguagem JAVA que percorre todos os termos bíblicos, excluindo-se as *stopwords*, e pesquisa pelo termo no Dicio de forma a retornar a estrutura da página contendo as informações de sinônimos, antônimos e verbos.

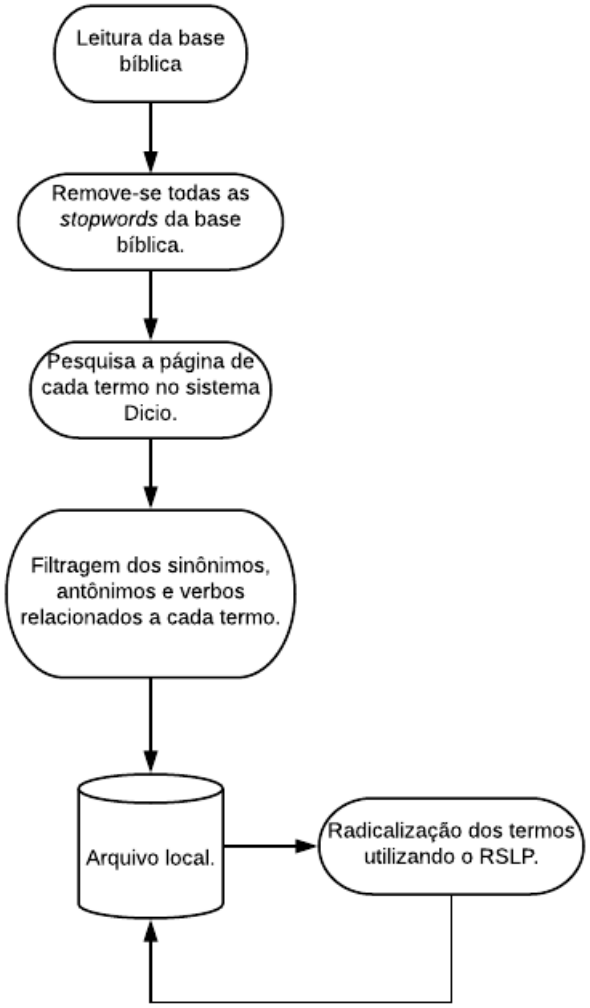
Com as páginas recuperadas, foi feito um processo de filtragem de informações, coletando os dados com expressões regulares para que os mesmos pudessem ser separados e organizados em um arquivo local, ordenados alfabeticamente.

Após as expressões sinônimas, antônimas e de flexão verbal estarem devidamente recuperadas e armazenadas, iniciou-se o processo de criação dos termos radicalizados, tanto os termos bíblicos quanto os termos recuperados pelo (7GRAUS, 2009-2016). Para tanto, utilizou-se o RSLP stemmer que é um projeto que foi aprimorado por Coelho (2007) e disponibilizado em linguagem C pelo mesmo. Portanto, para a radicalização dos termos utilizou-se o projeto em linguagem C, onde o mesmo recebia um argumento de parâmetro (que é o termo a ser radicalizado) e retornava uma *string* (termo radicalizado). Para o uso dessa ferramenta gerou-se um programa executável, de forma que o programa buscador idealizado em JAVA pudesse fazer a leitura dos termos salvos em arquivo e efetuasse a chamada do executável passando os termos em questão.

Com o programa executável devidamente gerado, o programa de busca iniciou a leitura de todos os termos gerados pelo arquivo, assim como os termos exatos da base bíblica, e chamou o executável salvando os termos radicalizados no mesmo arquivo, de forma a salvá-los em uma sequência chave-valor.

Com os passos gerados acima, deu-se a criação de um arquivo contendo todo o tesauro necessário para ser usado nesse projeto. Portanto, o arquivo continha tanto os termos exatos quanto os termos sinônimos, antônimos, flexões verbais e radicais. A figura 7 apresenta o fluxograma descrevendo esse processo.

Figura 7 – Fluxograma de criação do tesauro



Fonte: Autoria própria

4.2 Projeto do Banco de Dados

Esse projeto constitui-se de uma gama muito grande de termos isolados que devem fazer referência a um documento em específico. Tal explanação se dá pelo número de termos que esse projeto se propõe a trabalhar. Segue-se algumas características quantitativas do sistema.

- O sistema contém 228474 termos bíblicos no geral, onde apenas 28677 termos eram exatos na bíblia (excluindo-se repetições).

- O sistema contém 31097 versículos bíblicos, que foram considerados como os documentos dessa pesquisa.
- O sistema apresentou 64240 radicais distintos (excluindo-se repetições) considerando-se todos os termos usados.

De acordo com Silberschatz, Sundarshan e Korth (2016), os sistemas de banco de dados são projetados para gerir grandes massas de informação, assim como garantir a segurança das informações armazenadas. Com vista nessa informação, esse sistema se propôs a usar um banco de dados em vista a grande quantidade de informações a serem manipuladas, assim como a flexibilidade e rapidez com que essas informações pudessem ser recuperadas e exibidas aos avaliadores.

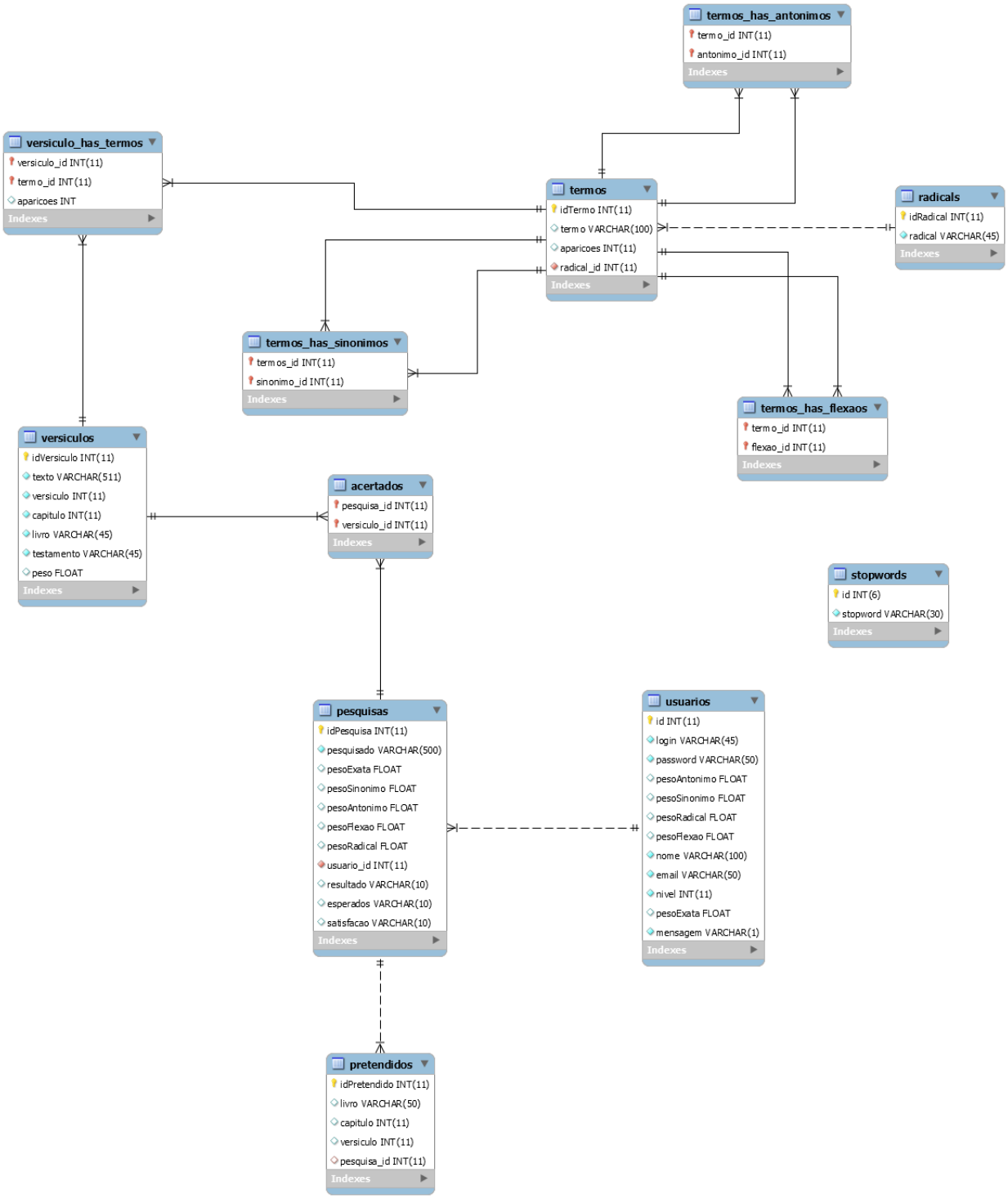
O banco foi projetado de forma a correlacionar o tesauro bíblico, ou seja, criar o relacionamento entre termos e documentos, assim como armazenar informações dos usuários, que são informações tanto pessoais quanto informações de avaliação do sistema. A figura 8 contém um desenho esquemático do banco de dados proposto para este trabalho.

Para gerar o banco, o programa de busca efetuou a leitura de todos os termos e suas derivações que estavam contidas no arquivo gerado pelo mesmo, e inseriu todos os termos em uma única tabela denominada termos. Essa tabela associa os termos de todos os contextos (termos_has_antonimos, termos_has_sinonimos e termos_has_flexaos) assim como uma ligação direta de todos os radicais pela tabela radicals, de forma a gerar a integração idealizada pelo tesauro. O programa também realizou a inclusão de todos os versículos bíblicos em uma única tabela, fazendo com que a mesma seja a coleção de documentos a que esse trabalho se propõe a analisar.

Para a avaliação do sistema, o banco contém a tabela de pesquisas, que armazena todas as pesquisas efetuadas pelos usuários, salvando os pesos dos contextos do usuário para cada pesquisa. Essa tabela contém um relacionamento direto com a tabela pretendidos, que são todos os versículos que o usuário considera relevante e que não foram retornados nos 10 primeiros resultados da busca, e com a tabela acertados, que são todos os versículos que foram retornados nos 10 primeiros resultados e que o usuário considerou relevante.

O banco foi acrescido de uma tabela com uma lista de *stopwords* do sistema, de forma que as buscas efetuadas tivessem *stopwords* removidas e ignoradas pelo sistema.

Figura 8 – Diagrama entidade relacionamento do banco de dados da aplicação



Fonte: Autoria própria

4.3 Implementação do Sistema de Avaliação

O sistema de avaliação constitui-se de parte importante deste trabalho, pois é com ele que haverá a interação direta do usuário com os dados do tesauro, além de ser responsável por salvar os dados de avaliação. Portanto, é de extrema importância que o mesmo seja de

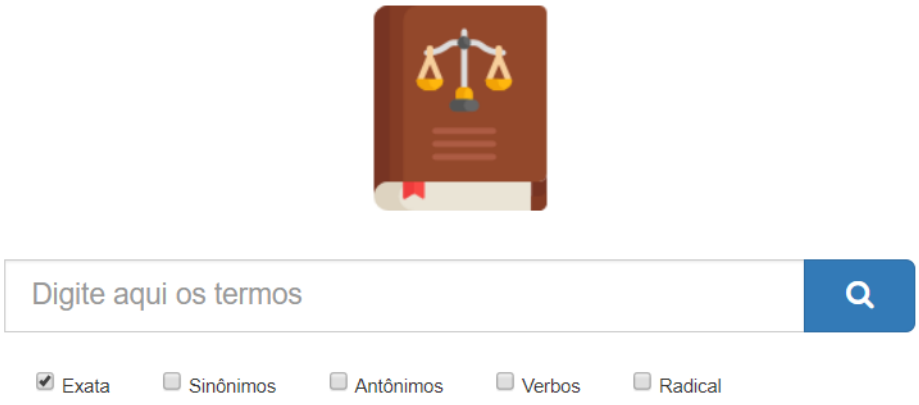
interface amigável e de boa usabilidade.

Para o *front-end* da aplicação, utilizou-se da linguagem de marcação HTML, CSS e da linguagem Javascript, usando-se na mesma a biblioteca Jquery. Junto ao conjunto apresentado utilizou-se do *framework* Bootstrap para que se fizesse o uso de componentes de interface pré-construídos.

Devido a baixa quantidade de telas a serem visualizadas pelos usuários e a simplicidade de interface da aplicação, neste trabalho optou-se pelo tipo de layout *One Page Layout*, de forma que as *views* fossem carregadas via AJAX Javascript em interação direta com o servidor.

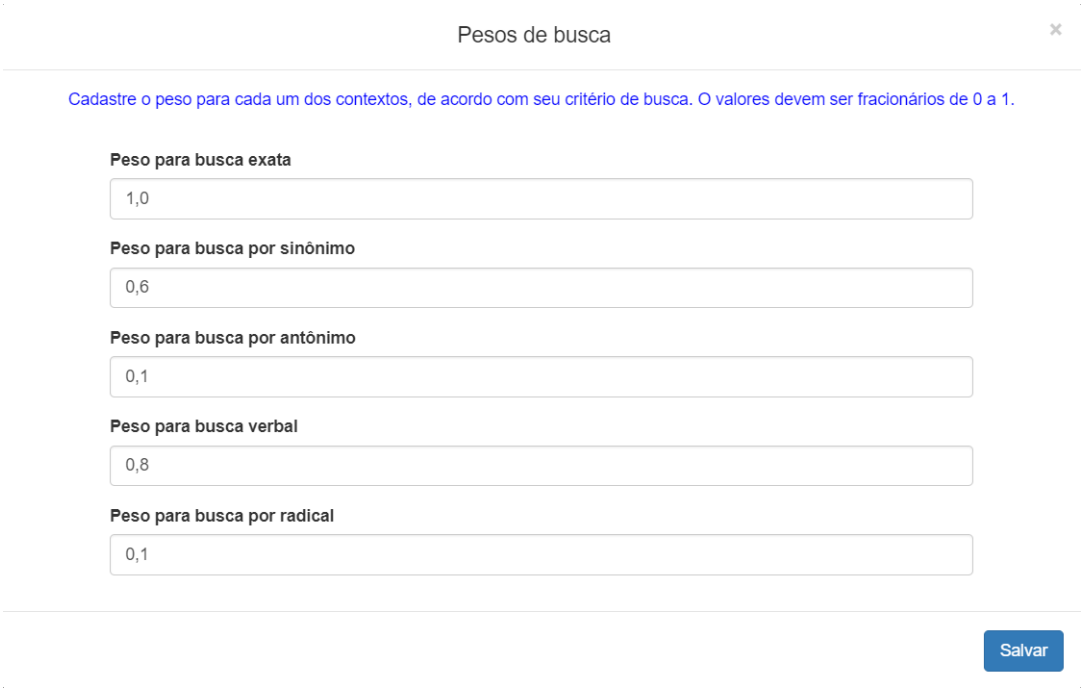
Nas figuras 9, 10, 11 e 12 mostram a interface do sistema em todos os passos executados pelo usuário para a avaliação desse sistema.

Figura 9 – Área de inserção da busca do usuário



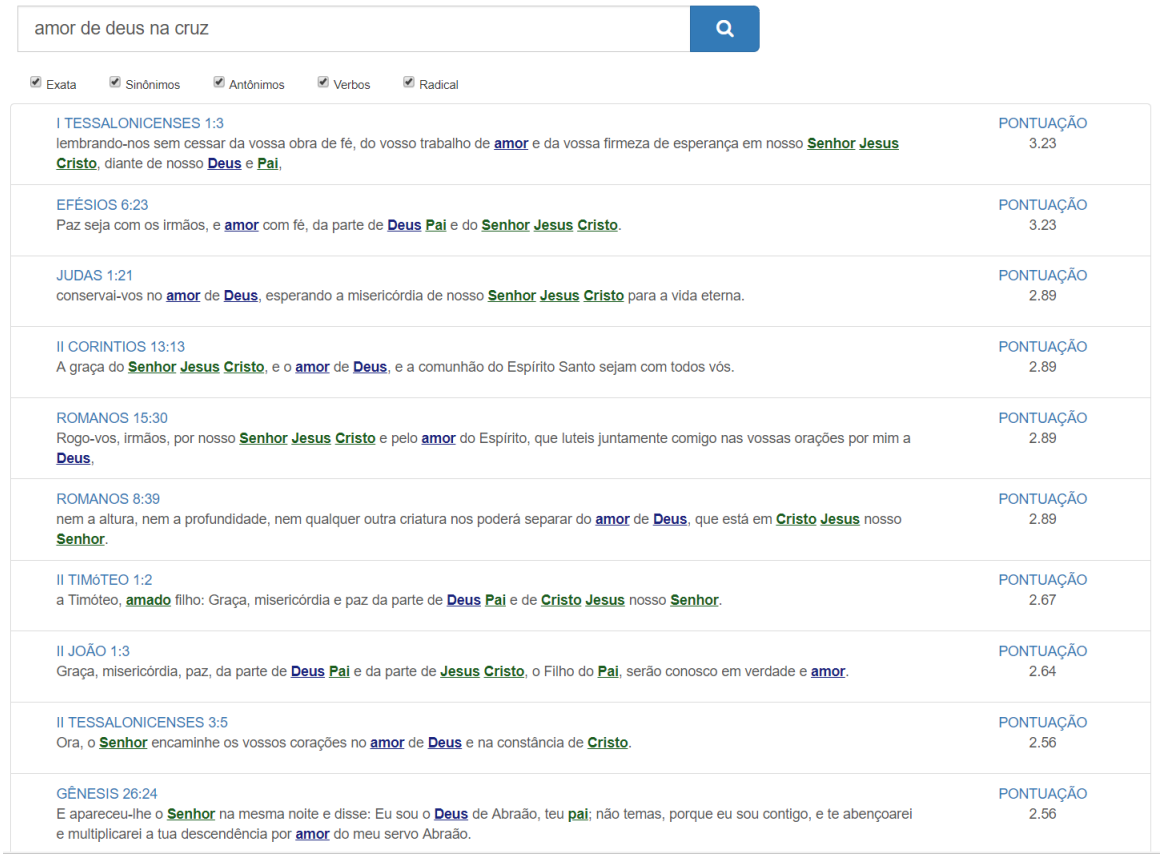
Fonte: Autoria própria

Figura 10 – Área de inserção de pesos dos contextos



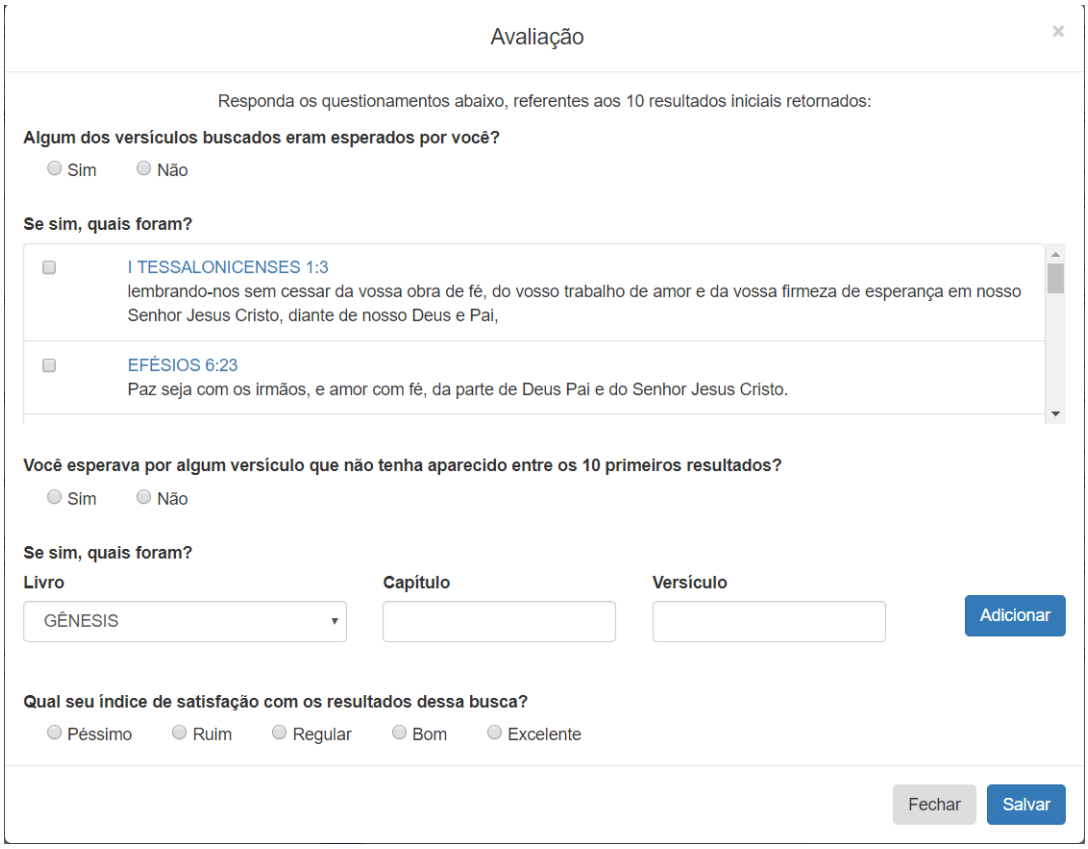
Fonte: Autoria própria

Figura 11 – Resultados da busca efetuada



Fonte: Autoria própria

Figura 12 – Questionário de avaliação dos usuários



Fonte: Autoria própria

Para o *back-end* da aplicação optou-se pela utilização da linguagem Ruby, utilizando-se do framework Ruby on Rails. Segundo Fuentes (2014), o framework foi pautado em cima das seguintes características:

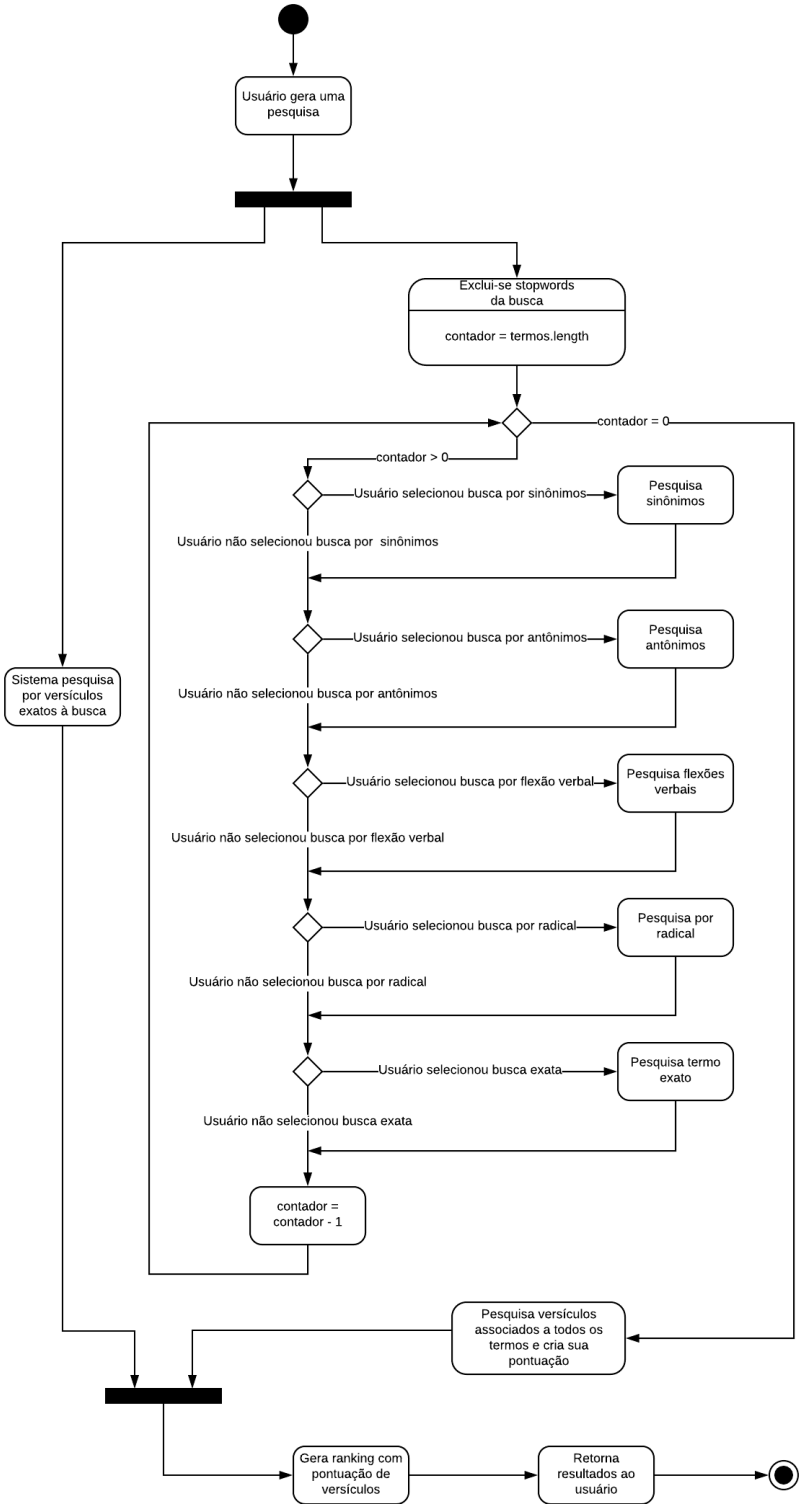
- "*Convention over configuration*", ou convenção a configuração, que é deixar de configurar uma série de estruturas em xml, e adotar uma convenção e alterá-la apenas quando necessário.
- "*Dont Repeat Yourself*", ou "não se repita", que significa que você nunca deve executar a mesma tarefa mais de uma vez.
- Automação de tarefas repetidas, que significa que um programador nunca pode perder tempo com tarefas repetidas, ocupando seu tempo sempre em tarefas interessantes.

Portanto, a utilização do Ruby on Rails se deu pela alta produtividade de código e praticidade no desenvolvimento, devido ao *framework* ter sido desenvolvido em cima de diretrizes da arquitetura MVC (*Model-View-Controller*) e ainda ser um projeto de código aberto com constantes melhorias pela comunidade de desenvolvimento.

A comunicação entre cliente-servidor nessa aplicação foi gerada de maneira híbrida, com partes da interação feita em requisições REST quando solicitadas via AJAX pelo lado do

cliente, ou utilizando-se de *partials* geradas pelo servidor em forma de conteúdo HTML quando solicitada em páginas estáticas. A figura 13 apresenta o fluxo de uso do sistema, para melhor compreendimento do funcionamento do mesmo.

Figura 13 – Diagrama de atividade UML do sistema de avaliação



Fonte: Autoria própria

O Ruby on Rails faz uso de um componente de acesso aos dados chamado de ActiveRecord. De acordo com Fuentes (2014), o ActiveRecord é um Object-Relational-Mapping que faz o mapeamento de estruturas relacionais a objetos Ruby. Portanto, é através dessa ferramenta que a aplicação faz a leitura direta do tesauro, assim como persiste as informações de avaliação gerada pelos usuários.

4.3.1 Ranqueamento das buscas

Neste trabalho optou-se pela utilização do modelo de RI clássico vetorial, que analisa o grau de similaridade entre os documentos e as consultas. Em específico nesse trabalho, os documentos de consulta são consideradas as buscas efetuadas pelo usuário, com a entrada de informações no sistema, e a coleção de documentos consiste em todos os versículos armazenados na tabela do banco de dados. Com base nisso, a pontuação resultante de cada versículo do sistema se dá com a proximidade de cada um dos mesmos diretamente com a pesquisa efetuada.

A proximidade das buscas com cada um dos versículos se dá também com a importância de um determinado contexto em relação a outros em específico a uma busca. Com base nisso, acrescentou-se ao grau de similaridade $sim(dj, q)$ o peso determinado por cada usuário a cada um dos contextos de forma a dar mais importância a versículos que estejam mais próximos a intenção dos usuário em relação a busca. Então o grau de similaridade proposto por este trabalho se dá por:

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

sendo que $W_{i,q}$ constitui-se do peso do contexto do termo em relação ao documento.

Portanto, levou-se em consideração não somente os termos exatos da pesquisa, mas sim todo o arcabouço de termos que possam estar relacionados aos termos de pesquisa do usuário. Logo, após a aplicação buscar por todos os termos correlacionados no banco de dados, fez-se então o ranqueamento de versículos passando cada um dos versículos pelo teste de ranqueamento descrito acima, criando uma pontuação a cada um dos mesmos. Através dessa pontuação, os versículos foram recuperados pelo grau de importância e retornados a página do usuário para análise dos mesmos.

4.3.2 Avaliação dos usuários

Uma parte importante desse trabalho constitui-se do *feedback* dos usuários. Esse *feedback* é importante para saber o quanto o sistema de fato satisfaz as intenções dos usuários em buscas pelo sistema.

A interface de avaliação constitui-se no preenchimento do questionário de satisfação do usuário, que corresponde aos 10 primeiros versículos retornados. A interface correspondente a esse formulário se encontra na figura 12. As questões a serem preenchidas pelo usuário eram:

- Algum dos versículos buscados eram esperados por você?
- Se sim, quais foram?
- Você esperava por algum versículo que não tenha aparecido entre os 10 primeiros resultados?
- Se sim, quais foram?
- Qual seu índice de satisfação com os resultados dessa busca?

Portanto, com base nas respostas dadas pelos mesmos constituiu-se os resultados desse trabalho descritos no capítulo 5.

5 Resultados

Este capítulo trata de criar uma análise dos resultados de avaliação do sistema por parte dos usuários, apresentando dados estatísticos para que através dos mesmos, seja feita uma conclusão crítica sobre o sistema.

5.1 Avaliadores

Os avaliadores selecionados para esse trabalho, constitui-se de um grupo de seis pessoas que se consideram ter o mínimo de conhecimento bíblico para essa avaliação. Ao se cadastrar no sistema, o usuário responde a pergunta sobre qual nível de conhecimento bíblico o mesmo considera ter, com as opções de alto, médio ou baixo. A tabela 1 apresenta a relação dos usuários com seu nível de conhecimento bíblico.

Tabela 1 – Relação de usuários com seu nível de conhecimento bíblico

Usuário	Nível de conhecimento bíblico
Usuário 1	Médio
Usuário 2	Médio
Usuário 3	Baixo
Usuário 4	Médio
Usuário 5	Alto
Usuário 6	Médio

Fonte: Autoria própria

Todas as avaliações foram feitas no mesmo notebook, sendo que o mesmo foi levado até os avaliadores para que eles estivessem mais confortáveis na avaliação do sistema.

Os avaliadores tiveram acesso direto ao sistema, sendo orientados apenas sobre a forma de uso do mesmo ou quando houve dúvidas sobre suas funcionalidades. As avaliações foram feitas sempre em local onde os usuários pudessem estar confortáveis e atentos para se avaliar o sistema.

Ao final da pesquisa todos os avaliadores assinaram um termo de consentimento livre e esclarecido, contendo as informações pela coleta de dados e sobre o propósito do trabalho.

Os usuários efetuaram quatro pesquisas no sistema, e cada pesquisa foi feita efetuando-se a alteração dos pesos dos contextos, sendo que a alteração de tais pesos foram feitas pelo próprio aplicador da avaliação do sistema. A tabela 2 mostra a relação dos pesos de cada contexto para cada uma das pesquisas.

Os pesos referentes a pesquisa 1 são os pesos padrão da aplicação, na hipótese de que as buscas exata e por sinônimo tem mais importância que as dos demais contextos. A pesquisa 2 oferece maior grau de importância ao peso do contexto radical, para avaliar se os resultados retornados por radical são relevantes aos usuários. A pesquisa 3 oferece menor

Tabela 2 – Relação de pesos dos contextos de acordo com as pesquisas

	Exata	Sinônimo	Antônimo	Verbo	Radical
Pesquisa 1	1	0,6	0,1	0,5	0,2
Pesquisa 2	0,1	0,1	0,1	0,1	0,8
Pesquisa 3	0,5	0,5	0,5	0,5	0,1
Pesquisa 4	0,5	0,5	0,5	0,5	0,5

Fonte: Autoria própria

grau de importância a busca por radical, para verificar se os resultados não radicalizados são mais importantes. A pesquisa 4 oferece o mesmo grau de importância a todos os contextos, para se verificar se todos os resultados contendo todos os contextos podem ser relevantes aos usuários.

Após os resultados serem mostrados aos usuários, os avaliadores marcaram quais versículos retornados foram importantes para aquela pesquisa, quais versículos eles consideravam relevantes e não foram retornados e um *feedback* de avaliação dos resultados retornados através de um grau de satisfação, sendo esses: péssimo, ruim, regular, bom e excelente.

Devido a alta correlação de tabelas presente no banco de dados da aplicação, houve uma alta média de tempo para as consultas efetuadas no banco, com o propósito de se gerar o ranqueamento de documentos. A tabela 3 apresenta alguns dos tempos relativos as pesquisa efetuadas na aplicação.

Tabela 3 – Tempo de resposta as consultas efetuadas na aplicação

Pesquisa efetuada	tempo de resposta(s)
vinha de nabote	15,86
Jesus ama o pecado	22,77
salvação de um rei	8,47
palavra revelada	8,83

Fonte: Autoria própria

5.2 Precisão e revocação

Os índices de precisão e revocação, são uma das métricas que geram uma análise qualitativa dos resultados gerados por um sistema de RI. Através desses índices tem-se uma base de análise para saber se os resultados retornados realmente produziram resultados relevantes para os usuários.

Segue-se abaixo a definição das medidas de precisão e revocação dada por Baeza-Yates e Ribeiro-Neto (2013):

- **Precisão** é a fração dos documentos recuperados que é relevante, isto é,

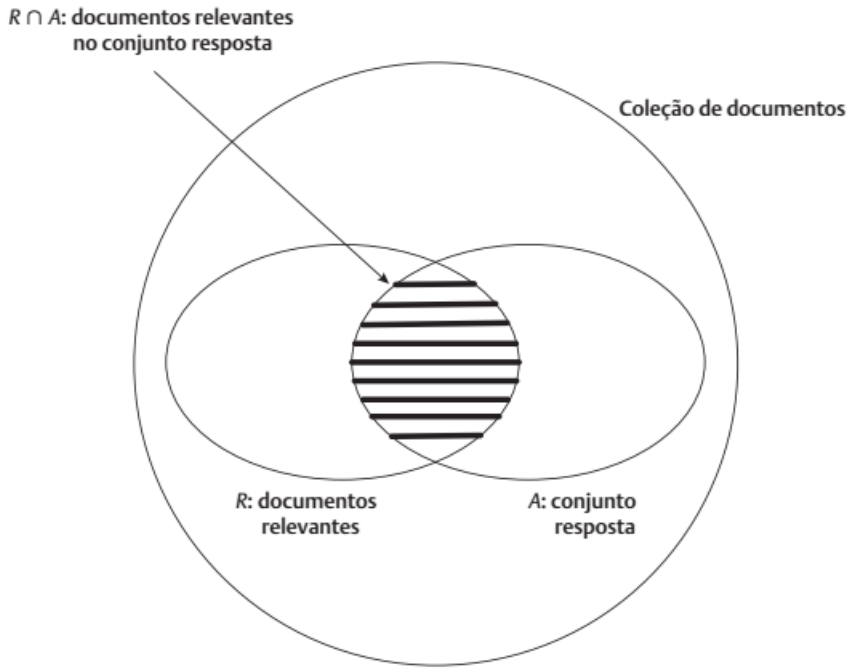
$$Precisão = p = \frac{|R \cap A|}{|A|}$$

- **Revocação** é a fração dos documentos relevantes que foi recuperada, isto é,

$$Revocação = r = \frac{|R \cap A|}{|R|}$$

R é a coleção de documentos relevantes recuperados e A é a coleção de documentos recuperados. A figura 14 exemplifica os valores de índices gerados de acordo com a coleção de documentos recuperados.

Figura 14 – Interseção dos conjuntos R e A



Fonte: (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 111)

Nesse trabalho os índices de precisão e revocação foram calculados com base nos primeiros dez versículos retornados para cada busca. Segundo Baeza-Yates e Ribeiro-Neto (2013), a precisão dos resultados da Web deve ser medida somente nas primeiras posições do ranking, podendo ser o conjunto dos 5, 10 e 20 primeiros resultados.

Foram considerados como o conjunto R os versículos que foram selecionados pelos usuários como relevantes, somados aos versículos que os usuários julgavam relevantes e não foram retornados considerando-se os primeiros dez resultados da busca, e o conjunto A os dez primeiros versículos retornados.

Analisando-se as pesquisas descritas na tabela 2, os índices de revocação e precisão de cada pesquisa estão descritas nas tabelas 4, 5, 6 e 7. Nelas estão descritas os índices gerados para as pesquisas de cada usuário, assim como a média geral de revocação e precisão.

As tabelas apresentam também o nível de satisfação do usuário com os resultados gerados. Os níveis de satisfação estão diretamente ligados a opção selecionada pelo usuário, sendo:

- Nível 1 - Péssimo
- Nível 2 - Ruim
- Nível 3 - Regular
- Nível 4 - Bom
- Nível 5 - Excelente

Tabela 4 – Índices de precisão, revocação e satisfação relacionados à Pesquisa 1

	Pesquisa	$\|R \cap A\|$	$\|R\|$	Precisão	Revocação	Satisfação
Usuário 1	vinha de nabote	9	9	0,9	1	5
Usuário 2	Deus ama o mundo	2	3	0,2	0,67	5
Usuário 3	senhor deus confiou	1	1	0,1	1	5
Usuário 4	a morte de jesus	1	1	0,1	1	5
Usuário 5	o bom pastor	2	2	0,2	1	5
Usuário 6	transformar água em vinho	2	2	0,2	1	4
Média	-	2,83	3	0,3	0,95	4,83

Fonte: Autoria própria

Tabela 5 – Índices de precisão, revocação e satisfação relacionados à Pesquisa 2

	Pesquisa	$\|R \cap A\|$	$\|R\|$	Precisão	Revocação	Satisfação
Usuário 1	carpintaria	0	2	0	0	1
Usuário 2	salvação de Nabucodonosor	0	0	0	0	2
Usuário 3	confiar	2	3	0,2	0,67	4
Usuário 4	jesus chorando	0	1	0	0	2
Usuário 5	viver pela fé	1	1	0,1	1	5
Usuário 6	Jesus ama o pecado	2	3	0,2	0,67	4
Média	-	0,83	1,67	0,08	0,39	3

Fonte: Autoria própria

Tabela 6 – Índices de precisão, revocação e satisfação relacionados à Pesquisa 3

	Pesquisa	$\|R \cap A\|$	$\|R\|$	Precisão	Revocação	Satisfação
Usuário 1	Viúva de Naim	1	2	0,1	0,5	3
Usuário 2	salvação de um rei	2	2	0,2	1	4
Usuário 3	salvação pela graça	0	1	0	0	2
Usuário 4	os deveres domésticos	0	1	0	0	2
Usuário 5	está consumado	1	1	0,1	1	5
Usuário 6	A fé que direciona o homem	2	3	0,2	0,67	5
Média	-	1	1,67	0,1	0,53	3,5

Fonte: Autoria própria

Tabela 7 – Índices de precisão, revocação e satisfação relacionados à Pesquisa 4

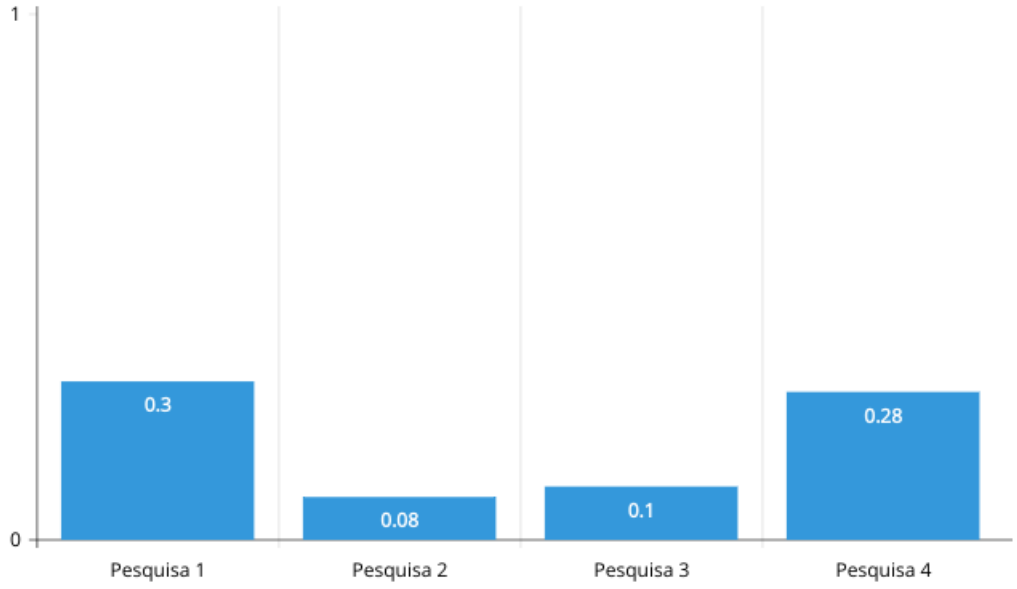
	Pesquisa	$\|R \cap A\|$	$\|R\|$	Precisão	Revocação	Satisfação
Usuário 1	cego no caminho	7	7	0,7	1	4
Usuário 2	salvação de uma mulher	0	0	0	0	2
Usuário 3	luz do mundo	3	3	0,3	1	4
Usuário 4	palavra revelada	4	5	0,4	0,8	5
Usuário 5	pela graça	1	1	0,1	1	5
Usuário 6	a palavra viva	2	3	0,2	0,67	5
Média	-	2,83	3,17	0,28	0,75	4,16

Fonte: Autoria própria

Conforme observado pelas tabelas apresentadas, os índices de precisão, revocação e satisfação médios foram maiores quando foi efetuada a Pesquisa 1, pois essa pesquisa utiliza-se dos pesos padrão da aplicação. Já na Pesquisa 2, os índices foram os menores devido a alta relevância dada ao peso por radical em detrimento aos outros. As pesquisas 3 e 4 apresentaram índices medianos, sendo o destaque feito a pesquisa 4 que apresentaram índices relativamente altos, quando comparados as pesquisas 1 e 2.

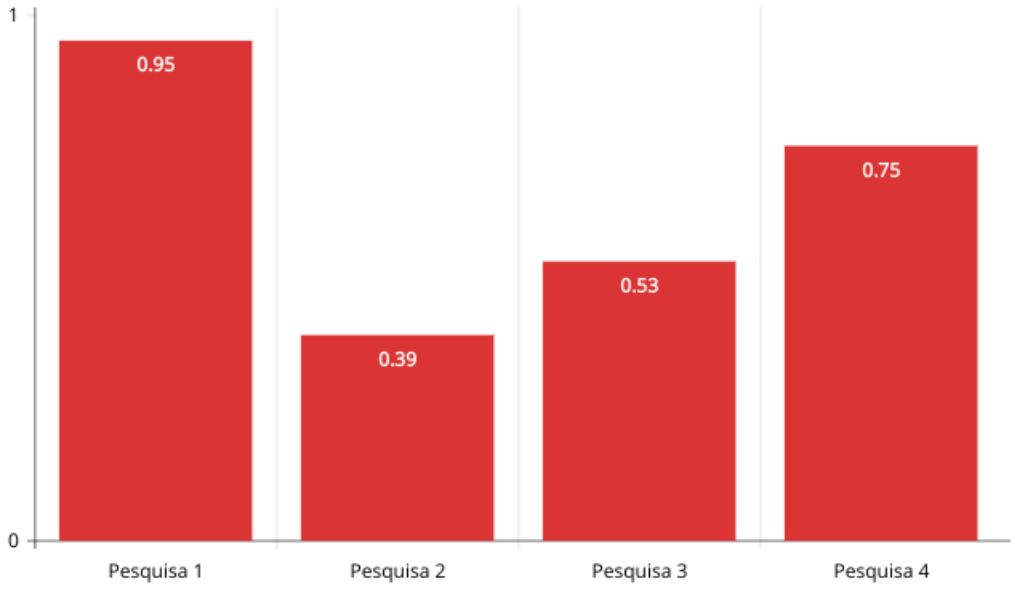
As figuras 15, 16 e 17 apresentam os gráficos de comparação entre os índices de cada pesquisa.

Figura 15 – Comparação dos índices de precisão médios entre as pesquisas



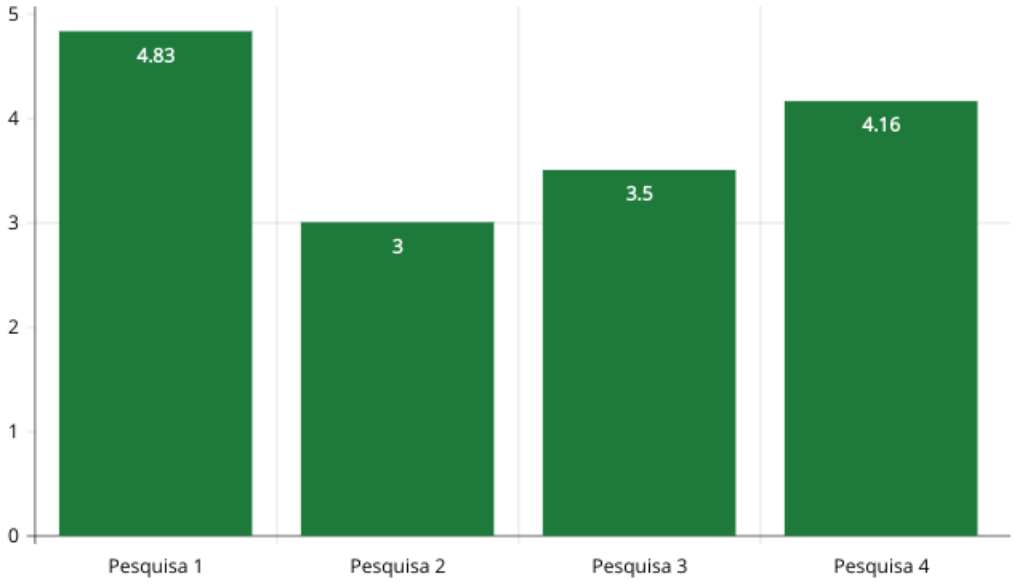
Fonte: Autoria própria

Figura 16 – Comparação dos índices de revocação médios entre as pesquisas



Fonte: Autoria própria

Figura 17 – Comparação dos índices de satisfação médios entre as pesquisas



Fonte: Autoria própria

6 Conclusão

Existem atualmente uma limitada quantidade de ferramentas que façam uma busca contextualizada sobre uma base complexa semanticamente como a bíblia sagrada. Mediante isso, neste trabalho foi implementado uma ferramenta que tem como proposta uma busca contextualizada sobre a bíblia. Para tal, utilizou-se da busca baseada em contextos específicos, utilizando-se de um tesauro genérico e radicalização de termos com o objetivo de retornar aos usuários resultados que possam ser relevantes a busca efetuada.

Com a implementação da ferramenta, foi feita a avaliação do sistema por um grupo de avaliadores detentores de conhecimento bíblico, visando medir os índices de precisão e revocação das buscas efetuadas. Apesar do uso de um banco de dados com alta performance como o MySQL, houveram alguns problemas relativos ao desempenho da aplicação, devido a alta correlação de termos. No entanto, a ferramenta manteve-se consistente para o uso dos avaliadores.

Através dos resultados avaliativos obtidos, pode-se observar que o índice de satisfação dos usuários e os índices de precisão e revocação foram maiores quando os pesos com valores padrão foram aplicados. Portanto, o uso de prioridades de contextos em um tesauro genérico se mostrou eficaz neste trabalho.

Por outro lado, o índice de satisfação dos usuários, de precisão e de revocação é menor quando o peso dos radicais foi maior, mostrando assim que apenas a pesquisa por radical pode não trazer resultados satisfatórios. Um fator que pode ter influenciado a isso é que a contextualização por radical pode relacionar termos que não tenham proximidade semântica.

Portanto, confirmou-se que o uso de um tesauro genérico aliado ao modelo de recuperação vetorial retornam resultados relevantes aos usuários, mesmo com o uso do contexto de termos radicalizados, devido aos altos índices de precisão e revocação média, alcançando assim os objetivos propostos pelo trabalho realizado.

Para trabalhos futuros, esse trabalho pode ser aprimorado no que diz respeito a melhoria de desempenho no ranqueamento de documentos, utilizando-se uma alternativa ao banco de dados relacional para se diminuir o tempo de recuperação de documentos da coleção. Também pode ser aprimorado no uso de informações dos usuários para melhoria no ranqueamento, atribuindo-se pesos aos documentos mais acessados pelos usuários. Pode-se também ser feita uma análise das técnicas aplicadas na bíblia sagrada em uma outra base específica, de forma a verificar a eficácia dos métodos aplicados neste trabalho em uma base distinta.

Referências

7GRAUS. *Dicio*. 2009–2016. Disponível em: <http://www.dicio.com.br/>. Acesso em: 12 ago. 2016. Citado nas páginas 13 e 26.

ANTUNES, E. J. B. Recuperação de informação em documentos jurídicos com expansão de consulta baseada em tesauro. Universidade Federal de Pernambuco, 2010. Citado na página 16.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca*. [S.l.]: Bookman Editora, 2013. Citado nas páginas 9, 10, 14, 15, 18, 19, 21, 22, 23, 24, 37 e 38.

BINDÁ, J. M.; BRANDT, M. A. G.; PIEDADE, M. P. Análise da aplicação de sistemas de recuperação de informação usando android numa base bíblica. 2013. Citado nas páginas 9 e 17.

CAMPOS, M. L. A.; GOMES, H. E. Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador. *Perspectivas em ciência da informação*, SciELO Brasil, v. 11, n. 3, p. 348–359, 2006. Citado na página 18.

COELHO, A. R. Stemming para a língua portuguesa: estudo, análise e melhoria do algoritmo rslp. 2007. Citado nas páginas 13, 19, 20 e 26.

EISBäR MEDIA GMBH. *Woxikon*. 2016. Disponível em: <http://sinonimos.woxikon.com.br/>. Acesso em: 12 ago. 2016. Citado na página 13.

FERNEDA, E. *Recuperação de informação: análise sobre a contribuição da ciência de computação para a ciência da informação*. 2003. Tese (Doutorado), 2003. Citado nas páginas 22, 23, 24 e 25.

FONSECA, J. da. *Apostila de metodologia da pesquisa científica*. João José Saraiva da Fonseca, 2002. Disponível em: <https://books.google.com.br/books?id=oB5x2SChpSEC>. Citado na página 12.

FUENTES, V. *Ruby on Rails: Coloque sua aplicação web nos trilhos*. [S.l.]: Casa do Código, 2014. ISBN 9788566250862. Citado nas páginas 32 e 34.

GERHARDT, T.; SILVEIRA, D. *Métodos de Pesquisa*. Plageder, 2009. (Série Educação a Distância - UFRGS). ISBN 9788538600718. Disponível em: <https://books.google.com.br/books?id=dRuzRyEIzmkC>. Citado na página 12.

GRANADA, R. L. Processos de construção automática de tesauro. Pontifícia Universidade Católica do Rio Grande do Sul, 2011. Citado nas páginas 18 e 19.

GREENBERG, J. Automatic query expansion via lexical–semantic relationships. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 52, n. 5, p. 402–415, 2001. Citado nas páginas 16 e 17.

KNAPP, S. *The Contemporary Thesaurus of Search Terms and Synonyms: A Guide for Natural Language Computer Searching*. Oryx Press, 2000. ISBN 9781573561075. Disponível em: <https://books.google.com.br/books?id=zYw3sYFtz9kC>. Citado na página 17.

MILANI, A. *MySQL - Guia do Programador*. [S.l.]: NOVATEC, 2007. ISBN 9788575221037. Citado na página 14.

MOURA, M. A. et al. Linguagens de indexação em contextos cinematográficos: a experiência de elaboração do tesauto eletrônico do cinema brasileiro. *Perspectivas em ciência da informação*, v. 10, n. 1, 2008. Citado na página 18.

ORENGO, V. M.; HUYCK, C. R. A stemming algorithmm for the portuguese language. In: *spire*. [S.l.: s.n.], 2001. v. 8, p. 186–193. Citado nas páginas 13, 20, 21 e 22.

PONTES, F. *Unofficial Python API for Dicio*. 2016. V1.0.0. Disponível em: <https://github.com/felipemfp/dicio>. Acesso em: 16 ago. 2016. Citado na página 13.

PONZETTO, S. P.; STRUBE, M. Deriving a large scale taxonomy from wikipedia. In: *AAAI*. [S.l.: s.n.], 2007. v. 7, p. 1440–1445. Citado na página 19.

PORTER, M. F. An algorithm for suffix stripping. *Program*, MCB UP Ltd, v. 14, n. 3, p. 130–137, 1980. Citado na página 20.

ROGET, P. M. *Roget's Thesaurus of English Words and Phrases...* [S.l.]: TY Crowell Company, 1911. Citado na página 18.

SILBERSCHATZ, A.; SUNDARSHAN, S.; KORTH, H. *Sistema de Banco de Dados*. [S.l.]: Elsevier Editora Ltda., 2016. ISBN 9788535251425. Citado na página 28.

VOORHEES, E. M. On expanding query vectors with lexically related words. In: *TREC*. [S.l.: s.n.], 1993. p. 223–232. Citado na página 17.

XAVIER, B. M.; GOMES, G.; SILVA, A. Análise comparativa de algoritmos de redução de radicais e sua importância para a mineração de texto. *Pesquisa Operacional para o Desen-volvimento*, v. 5, n. 1, p. 84–99, 2013. Citado na página 13.

ZIRKELBACH, A. *Sinonimos-Online*. 2012–2016. Disponível em: <http://sinonimos-online.com/>. Acesso em: 12 ago. 2016. Citado na página 13.