Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

O uso do TeP 2.0 em sistemas de recuperação de informação

Maria Rosângela de Almeida

JUIZ DE FORA
JULHO, 2016

O uso do TeP 2.0 em sistemas de recuperação de informação

Maria Rosângela de Almeida

Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Orientador: Tarcísio de Souza Lima

JUIZ DE FORA
JULHO, 2016

O uso do TeP 2.0 em sistemas de recuperação de informação

Maria Rosângela de Almeida

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Tarcísio de Souza Lima
M. Sc. em Informática, PUC-Rio

Regina Maria Maciel Braga
D.Sc. em Engenharia de Sistemas e Computação, UFRJ

Victor Ströele de Andrade Menezes
D.S. em Engenharia de Sistemas e Computação, UFRJ

JUIZ DE FORA
28 DE JULHO, 2016

À minha mãe pelo amor incondicional.

Resumo

O mundo vive em um tempo em que a informação é um bem precioso e esse bem precisa estar disponível de forma clara, direta e em tempo hábil ao usuário. Assim, é fundamental o constante aprimoramento dos sistemas de recuperação da informação. Porém essa não é uma tarefa trivial, já que na maioria desses sistemas o usuário digita um pequeno conjunto de palavras e espera que o conteúdo retornado seja exatamente aquilo que procurava. O objetivo deste trabalho é fazer uma análise do impacto do uso do thesaurus TeP 2.0 em um sistema de recuperação da informação verificando seu desempenho ante as métricas mais comuns utilizadas na área. Assim, como resultado, será desenvolvida uma ferramenta de busca de documentos para pesquisa na base de atas do Conselho Superior da UFJF.

Palavras-chave: Recuperação de Informação, *thesaurus*, TeP 2.0

Agradecimentos

A Deus por me iluminar todos os dias.

À minha mãe pelo apoio e ao meu pai por custear meus estudos. Ao meu namorado pela compreensão.

Aos professores do Departamento de Ciência da Computação, em especial ao professor Tarcísio de Souza Lima pela orientação e aos professores Regina Maria Maciel Braga e Jairo Francisco de Souza pelo suporte neste trabalho. A todos os funcionários do Instituto de Ciências Exatas pelo carinho. A todos os amigos que dividiram comigo esta experiência.

"O sucesso é ir de fracasso em fracasso sem perder entusiasmo."

Winston Churchill

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

UFJF	Universidade Federal de Juiz de Fora
RI	Recuperação de Informação
TF	Term Frequency
IDF	Inverse Document Frequency
RSLP	Removedor de Sufixos para a Língua Portuguesa
TeP	Thesaurus para o português do Brasil

1 Introdução

1.1 Apresentação do Tema

O conceito informação é bastante amplo, pode se referir ao que é escrito exatamente em palavras ou possuir um sentido conotativo em que existe margem para diversas interpretações. Em uma sociedade portadora de tanta informação não falta detalhamento sobre os diversos assuntos que alguém possa ter interesse, o grande problema é fazer a localização dessa informação. Assim, os sistemas de recuperação da informação objetivam fornecer ao usuário a informação de forma precisa e em tempo hábil.

A pesquisa em um texto é um tema bastante discutido e trabalhado, porém, fazer uma busca que tenha como retorno a informação que realmente atenda às necessidades e objetivo de quem solicita pode não ser uma tarefa fácil. Cada vez mais os sistemas de recuperação da informação vêm aprimorando-se nesse sentido. No Conselho Superior (CONSU) da UFJF isso não é diferente. As atas de suas reuniões são longas e atualmente armazenadas em arquivos de texto, o que dificulta a localização de determinadas informações.

1.2 Justificativa

Segundo seu próprio regimento, o Conselho Superior(CONSU) é o órgão máximo de deliberação interna da Instituição, com definição estatutária, possuindo função normativa, deliberativa e de planejamento da UFJF. Ele possui grandes responsabilidades na gestão de assuntos da UFJF e em suas reuniões são discutidos diversos temas de interesse da universidade. Durante os encontros são redigidas atas sobre decisões tomadas e posteriormente fica bastante complicado o acesso a essas informações sem que haja um sistema de busca eficiente. Será muito útil para a sociedade poder acompanhar as decisões do CONSU e também para os membros do conselho acessarem os temas de interesse de maneira mais precisa.

1.3 Objetivos

O objetivo principal deste trabalho é fazer um estudo crítico sobre o uso do *thesaurus* TeP 2.0 nos sistemas de recuperação de informação, em especial, no desenvolvimento de uma ferramenta de pesquisa baseada na coleção de atas de CONSU.

Como objetivos específicos deste trabalho destaca-se:

- Realizar análise crítica de processos de stemização;
- Testar comportamento de pesquisas com utilização de um *thesaurus*;
- Realizar análise crítica de processos de indexação.

1.4 Organização

No próximo capítulo serão apresentados tópicos relevantes sobre fases que são executadas no processo de recuperação da informação. Este capítulo propõem uma fundamentação teórica sobre o tema e serve de base para a execução de um projeto que será desenvolvido.

No terceiro capítulo serão apresentadas as fases de desenvolvimento do projeto de recuperação da informação nas atas do CONSU. Será realizado de acordo com o encaminhamento proposto no capítulo de fundamentação teórica.

No quarto capítulo serão expostos os resultados práticos atingidos pelo projeto. Esses resultados serão baseados nas métricas de revocação e precisão, comuns para avaliação na área de recuperação da informação.

No quinto capítulo serão discutidas as possíveis vantagens ou desvantagens da utilização do TEP 2.0 no processo de recuperação das atas do CONSU segundo as métricas aplicadas no capítulo de resultados.

2 Fundamentação teórica

Em uma pesquisa a fundamentação teorica é essencial ao atendimento da qualidade do trabalho e seu objetivo é validar as conclusões propostas. Deve partir da análise rigorosa dos dados e propor um diagnóstico preciso ao enfrentamento do problema apresentado. Além de ajudar o pesquisador a traçar um plano de ações e uma estratégia de avaliação de resultados.

2.1 Indexação de Documentos

Normalmente, os sistemas de RI utilizam indexadores para facilitar o trabalho dos algoritmos de consulta. A técnica de mapeamento de palavras é mais utilizada em grandes coleções de documentos, podendo armazenar todas as palavras do arquivos ou apenas termos que melhor o descrevem. Entre as diversas estruturas de indexação, as listas invertidas são as mais utilizadas, pela simplicidade de sua estrutura, eficiência nas buscas, adequação a vários tipos de granularidade do índice e possibilidade de compressão (Neubert, 2000).

2.1.1 Listas Invertidas

Uma lista invertida é uma estrutura de dados capaz de armazenar os distintos termos de uma coleção de documentos e um identificador de cada documento ao qual pertence essa palavra. Para Corrales (2005), a lista invertida é similar a uma base de dados relacional, porém em um nível de abstração mais baixo. Para exemplificar, a tabela 2.1 apresenta uma coleção com quatro documentos e seus conteúdos. Entre as diversas formas de construção das lista pode-se usar o modelo apresentado na tabela 2.2 ou na tabela 2.3.

Tabela 2.1: Exemplo de uma coleção textual

Documento	Texto
1	A computação é uma arte.
2	A computação é amada.
3	Toda arte é bela e é amada.

Tabela 2.2: Exemplo de uma lista invertida que utiliza frequência

Vocabulário	(Documento;Frequência)
A	(1;1)(2;1)
amada	(2;1)(3;1)
arte	(1;1)(3;1)
bela	(3;1)
computação	(1;1)(2;1)
e	(3;1)
é	(1;1)(2;1)(3;2)
toda	(3;1)
uma	(1;1)

Tabela 2.3: Exemplo de uma lista invertida que utiliza posição

Vocabulário	(Documento;Posição)
A	(1;1)(2;1)
amada	(2;4)(3;7)
arte	(1;5)(3;2)
bela	(3;4)
computação	(1;2)(2;2)
e	(3;5)
é	(1;3)(2;3)(3;3)(3;6)
toda	(3;1)
uma	(1;4)

A tabela 2.2 armazena o vocabulário e o número de vezes que o termo aparece em cada documento da coleção, já a tabela 2.3 se preocupa em indicar a posição exata de ocorrência da palavra o que permite, inclusive, o cálculo da frequência apresentada na tabela 2.2.

Em listas invertivas estendidas mapeiam-se também a posição de ocorrência e a frequência em que o termo aparece no documento. Armazenar esses dados favorece o desenvolvimento de algoritmos de busca que façam ordenação por relevância baseados na frequência e proximidade dos termos de interesse. A utilização desses arquivos indexadores visa retornar resultados de forma rápida e eficiente uma vez que, utilizando algorítmos adequados, pode-se chegar diretamente ao termo pesquisado.

Os pontos negativos de seu uso ficam por conta da necessidade constante de atualização das listas quando um documento for inserido ou retirado da coleção e, além disso, pela busca que é, normalmente, feita exatamente pelos termos digitados pelo usuário o que exclui da relação aqueles documentos que não possuem as palavras buscadas mas

que podem ser relevantes para a consulta como termos sinônimos.

Além dos modelos de arquivos apresentados nas tabelas 2.2 e 2.3, existem outras abordagens possíveis para a criação de listas invertidas, inclusive com eliminação de parte do vocabulário e a utilização prioritária de palavras substantivas. Adjetivos, advérbios e conectores são menos úteis como termos de indexação pois funcionam principalmente como complementos (Baeza-Yates; Ribeiro-Neto, 2011).

2.2 Os Modelos de Recuperação de Informação

Modelos de recuperação da informação são padrões definidos que visam retornar a informação solicitada pelo usuário basicamente através do uso de palavras-chave. Entre os diversos modelos de recuperação de informação, são considerados modelos clássicos: o Modelo Booleano, o Modelo Vetorial e o Modelo Probabilístico. Para Ferneda (2003) os sistemas de recuperação da informação devem representar o conteúdo dos documentos do *corpus* e apresentá-lo ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfazem total ou parcialmente à sua necessidade de informação.

Muitos dos modelos de recuperação da informação foram desenvolvidos nas décadas de 60 e 70; ainda assim seus fundamentos continuam sendo usados nos dias atuais. A qualidade da aplicação desses modelos está relacionada ao modelo escolhido diante de um contexto de aplicação e utilização. Assim, a eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que ele utiliza e é influenciada por seu modo de operação. Assim, escolhido bem o modelo mais adequado, aproxima-se do grande objetivo que é entregar ao usuário as informações relevantes à sua pesquisa.

2.2.1 Modelo Booleano

Baseado na Teoria dos Conjuntos e na Álgebra de Boole, esse modelo se apresenta como um dos mais simples e fáceis de implementar. Suas consultas se baseiam na existência ou não dos termos buscados e o resultado é representado pelos documentos que satisfaçam as relações estabelecidas logicamente, ocasionando assim uma falta de meio termo entre documentos que estão diretamente ligados ao retorno e documentos não relacionados. O

Modelo Booleano prevê que cada documento seja relevante ou não relevante. Não existe satisfação parcial das condições da consulta (Baeza-Yates; Ribeiro-Neto, 2011).

O tipo de pesquisa utilizado é bastante formal e faz uso dos operadores AND, OR e NOT para exprimir as relações desejadas entre termos dos documentos, conforme a figura 2.1. Se por um lado garante formalidade e intuitividade na pesquisa ao usuário qualificado, para o usuário leigo pode ser bastante complexo o entendimento das relações booleanas, podendo esse não chegar ao retorno desejado ou pior direcioná-lo a informação errada ou fazê-lo acreditar que o conteúdo buscado não existe.

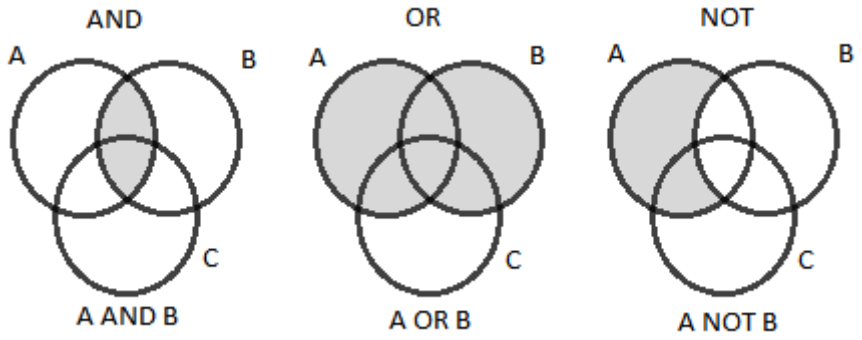

Figura 2.1: Representação dos conectivos utilizados no Modelo Booleano

A Álgebra Booleana confere um carácter binário ao retorno da pesquisa realizada pelo usuário, cada documento da coleção pertence ou não ao conjunto de documentos relevantes. Esse critério binário de decisão, sem nenhuma noção de grau, impede uma boa qualidade na recuperação (Baeza-Yates; Ribeiro-Neto, 2011). Outro grande obstáculo de sua utilização é a inexistência de mecanismo pelo qual os documentos resultantes da busca possam ser ordenados por ordem de relevância, o que dificulta o usuário encontrar o documento desejado entre os vários documentos retornados. De forma análoga os termos indexados são vistos com a mesma valoração; sendo assim, nenhum termo é visto como o melhor descritor do documento. Mesmo com esses entraves, o modelo booleano está presente em muitos sistemas de recuperação de informação dada sua facilidade de implementação.

2.2.2 Modelo Vetorial

Verifica-se que o Modelo Booleano é bastante ineficiente quando o objetivo é uma busca em que se deseja um ranqueamento de retornos pela sua relevância. Assim, o Modelo Vetorial sugere uma maneira em que cada termo receba um tratamento especial diante da sua importância na representação do documento.

Para ajustar melhor esse retorno, de maneira independente, cada palavra utilizada na busca recebe um peso não binário e fica evidenciado que o importante é saber o peso que cada termo da consulta possui para descrever um documento. No modelo vetorial um documento é representado por um vetor onde cada elemento representa o peso, ou a relevância, do respectivo termo de indexação para o documento (Ferneda, 2003). Um exemplo de representação vetorial é apresentado na figura 2.2.

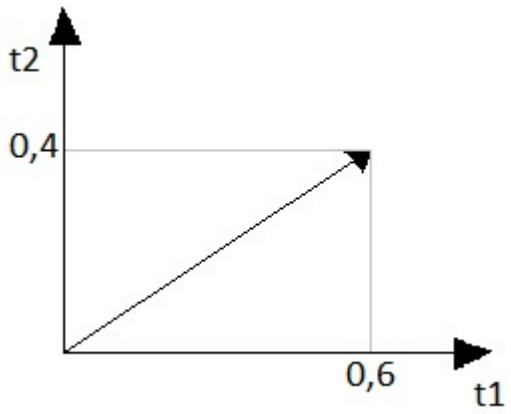

Figura 2.2: Exemplo de representação do modelo vetorial com dois termos indexados

Segundo Baeza-Yates; Ribeiro-Neto (2011),

> para o Modelo Vetorial, o peso w_{ij} associado ao par termo-documento (k_i, d_j) é não negativo e não binário. Os termos de indexação são todos considerados mutuamente independentes e são representados por vetores unitários e em um espaço com t dimensões, no qual t é o número de termos de indexação. As representações do documento d_j e da consulta q são vetores com t dimensões dadas por
>
> $$\vec{d_j} = (w_{1,j}, w_{2,j}, ...w_{t,j},) \tag{2.1}$$
>
> $$\vec{q} = (w_{1,j}, w_{2,j}, ...w_{t,j},) \tag{2.2}$$
>
> onde w_{iq} é o peso associado ao par termo-consulta $(k_{i,q})$, com $w_{i,q} \geq 0$.

Assim, todas as consultas e documentos podem ser representados em um espaço multidimensional em que a distância entre os vetores exprime a similaridade entre os

documentos da coleção. Dessa forma, o resultado da busca é uma relação ordenada de documentos segundo um cálculo específico de relevância.

As principais vantagens do Modelo Vetorial são sua simplicidade de implementação e sua eficiência diante das mais diversas coleções de documentos. Segundo Ferneda (2003), o modelo vetorial permite o desenvolvimento de soluções simples e rápidas e, por essas razões, o modelo vetorial é amplamente utilizado em soluções para indexação e pesquisa de documentos. Já uma das possíveis desvantagens que pode-se apontar para esse modelo é a condição dos termos de indexação, que são totalmente independentes. No entanto, não se pode garantir que essa seja uma desvantagem já que uma utilização indiscriminada das relações de posição entre os termos pode atrapalhar o retorno desejado pelo modelo.

2.3 Ponderação de Termos

Na indexação de termos dos documentos pode-se observar que cada termo tem um grau de importância para representar um documento. Assim, por exemplo, um termo que aparece uma vez na coleção é muito importante para descrever uma busca, ao contrário de um termo que aparece na maioria dos documentos. Termos que aparecem mais vezes em um documento também podem ser considerados importantes no ranqueamento de resultados em um sistema de busca. Esses padrões que tentam qualificar a importância de termos podem ser analisados de maneira conjunta, o que confere melhor ordenação ao resultado da busca.

A frequência do termo (TF, *Term Frequency*) e a frequência inversa de documento (IDF, *Inverse Document Frequency*) são fundamentos do esquema de ponderação mais popular em RI, chamado TF-IDF, que iremos discutir agora (Baeza-Yates; Ribeiro-Neto, 2011).

2.3.1 Ponderação da Frequência dos Termos (TF)

Nesse esquema de ponderação, o peso do termo t_i é proporcional ao número de vezes (frequência) f_{ij} que o termo ocorre no documento d_j. Normalmente, em sistemas mais simples, utiliza-se valor binário. Assim a frequência do termo k_i em um documento d_j

pode ser definida conforme a equação 2.3.

$$tf_{ij} = f_{ij} \tag{2.3}$$

Como usualmente a ponderação da frequência dos termos é utilizada em conjunto com a ponderação da frequência inversa dos documentos, pode-se utilizar uma variação logarítmica da TF para que essa possa ser compatível com a IDF, que será tratada no próximo tópico. Assim, chega-se à equação 2.4.

$$tf_{ij} = 1 + \log f_{ij} \tag{2.4}$$

2.3.2 Ponderação da Frequência Inversa dos Documentos (IDF)

O principal motivo do uso desse fator IDF se deve ao fato de que uma palavra que apareça em muitos documentos não pode ser usada para distinguir os objetos da coleção (Tsuji; Kamaura, 2008). Assim, a utilização desse fator de ponderação garante que termos de alta frequência no documento tenham seu peso diminuído e termos menos frequentes sejam considerados mais relevantes para a consulta. Para calcular o valor IDF_i de cada termo da coleção pode-se usar a equação 2.5

$$IDF_i = \log \frac{N}{n_i} \tag{2.5}$$

onde N é o número de documentos da coleção e n_i é a frequência de um termo na coleção.

2.3.3 Ponderação TF-IDF

A utilização da ponderação da frequência dos termos e a ponderação da frequência inversa dos documentos podem ser combinadas de acordo com a equação 2.6 para calcular o peso w_{ij} do termo k_i no documento d_j,

$$w_{ij} = (1 + \log f_{ij}) \times \log \frac{N}{n_i} \tag{2.6}$$

Para (Baeza-Yates; Ribeiro-Neto, 2011), embora simples, os pesos TF-IDF são

bastante eficazes para coleções genéricas, isto é, para atribuir pesos aos termos de uma coleção de documentos sobre a qual não temos nenhuma informação.

2.4 O Uso de *Stopwords*

Em um sistema de recuperação de informação nem todas as palavras contidas nos documentos são boas descritoras para catalogá-los. Termos de ligação e palavras que aparecem com alta frequência na coleção podem prejudicar o ranqueamento dos resultados das buscas. Para exemplificar, pode-se imaginar uma coleção de documentos históricos brasileiros, a palavra "Brasil" deve ocorrer em boa parte deles e não é boa localizadora de um documento específico. Normalmente, pronomes, artigos, preposições, conjunções e verbos de ligação não possuem relevância na indexação de documentos e são chamados de *stopwords*. Sua remoção da lista de índices busca melhorar o ranqueamento e a classificação dos documentos e, além disso, tornar as consultas mais eficientes e rápidas. Segundo Arantes (2005), essas palavras não são capazes de discriminar documentos e também não devem constar na estrutura de índice.

Em geral, a definição de um termo ser considerado ou não uma *stopword* parte de uma lista preestabelecida de palavras mais comuns do idioma que pode ser aplicada sobre coleções genéricas nos sistemas de RI. De fato, essas listas podem ser bastante úteis, em especial, em coleções pouco dinâmicas cujos documentos são pouco alterados e não há inserção frequente de novos itens. No entanto, a lista de *stopwords* pode ser deduzida diretamente do sistema de RI, voltada para a coleção que será aplicada, a partir do processamento e identificação dos termos mais frequentes nos documentos. Isso varia de acordo com os temas abordados na coleção, mas garante cobertura integral das palavras menos relevantes dentro desse contexto.

Pode-se definir qualquer lista de *stopwords*, mas esta escolha impacta diretamente na qualidade do retorno da busca realizada pelo usuário. Inclusive, no caso de pesquisa por frases completas, é possível que sua utilização seja dispensada já que pode trazer anomalias aos resultados. Por exemplo, na busca da sequência "Banco de Dados"o termo "de"provavelmente será eliminado. Essa operação também pode ser considerada uma técnica de compressão de textos, pois a eliminação de *stopwords* reduz o número de

palavras a serem analisadas no documento e também o número de palavras a serem armazenadas em uma base de dados. (Dias; Malheiros, 2006)

2.5 Stemização

Um termo de mesma relevância para o sistema de recuperação da informação pode apresentar-se de maneira distinta entre a busca realizada pelo usuário e as palavras que constam nos documentos da coleção. Palavras flexionadas em gênero e número ou que apresentam sufixos podem ter grande importância para que sejam recuperadas as informações desejadas. No processo de indexação os termos são relacionados como se apresentam no texto, mas uma busca considerando apenas termos exatamente iguais aos listados tende a ser bastante ineficiente. Um usuário que pesquisa por "estrela"provavelmente interessa-se também por trechos que contenham o termo "estrelas". Assim, uma busca que não considere as palavras de acordo com seu radical não recupera termos que apresentam semânticas diretamente relacionadas.

Segundo Oliveira (2015), as etapas de pré-processamento adotadas visam diminuir a complexidade do vocabulário de termos considerados para treinamento e estabelecer um meio de extrair o potencial completo do conjunto de dados. O que torna necessário um processo de stemização dos termos indexados.

A stemização é um procedimento de tratamento nas palavras que faz sua redução ao radical por meio da retirada de sufixos e prefixos apresentando-as na sua forma básica, independentemente de sua categoria gramatical. Sua raiz deve ser selecionada de modo que seja suficiente para relacionar-se a outros termos que estão em um mesmo contexto morfológico. De acordo com a Wikipédia (2016), vários motores de buscas tratam palavras com o mesmo tronco como sinônimos como um tipo de expansão de consulta, em um processo de combinação.

O processo de stemização é utilizado principalmente para melhorar algoritmos de recuperação da informação na expansão de consulta tornando-os mais eficientes, mas pode também ajudar a diminuir o número de termos no processo de indexação, já que o uso de um radical contempla diversas palavras dentro de um contexto morfológico.

Devido à grande quantidade de variações e flexões sofridas pelas palavras no

português, existe muita dificuldade na criação de um algoritmo de stemização que apresente poucas incertezas quanto à qualidade dos radicais retornados. Uma proposta a esse problema e que apresenta resultados satisfatórios é o algoritmo RSLP Stemmer.

2.6 RSLP Stemmer (Removedor de Sufixos para a Língua Portuguesa)

Segundo Longhi et al (2009), a remoção de prefixos pode resultar numa antonimação da palavra, remetendo a um resultado inverso ao expressado. Assim o algoritmo RSLP Stemmer estabelece uma série de regras para retirar o sufixo dos termos que serão indexados. Essas regras obedecem o padrão: sufixo; tamanho mínimo do radical; *string* de substituição; lista de exceções. Considerando o seguinte exemplo de regra:

{"gue",2,"g",{"gangue","jegue"}},

o sufixo "gue" é removido no processo de stemização. O número "2" é o tamanho mínimo que o radical pode ter após a retirada do sufixo. A letra "g" será incorporada ao radical que tiver seu sufixo extraído por essa regra. As palavras "gangue" e "jegue" são exceções e não devem sofrer a extração de sufixo.

O RSLP Stemmer proposto por Orengo; Huyck (2001) e implementado originalmente em linguagem C, possui oito etapas que devem ser seguidas na seguinte ordem:

- Etapa 1: Redução de Plural

  Na maioria dos casos as palavras em português que se apresentam no plural são terminadas por "s", mas nem todas terminadas em "s" encontram-se no plural. Nessa etapa faz-se a remoção do "s" mas com algumas substituições, por exemplo "anéis", retira o "éis" e troca por "el". As palavras que terminam em "s" e não são plurais devem ser tratadas como exceções, por exemplo "país".

- Etapa 2: Redução de Feminino

  Os substantivos e adjetivos no português possuem classificação de gênero. Essa etapa busca transformar palavras femininas em masculinas. Para isso apresenta os sufixos mais comuns femininos e seus correspondentes masculinos, por exemplo

"senhora" perde o sufixo "ora" e ganha o "or".

- Etapa 3: Redução de Advérbio

  Esta etapa é bastante direta e possui uma única regra que é a eliminação do sufixo "mente", mas como nem todas palavras do português que possuem essa terminação são advébios, por exemplo "experimente", é necessária uma lista de exceções.

- Etapa 4: Redução de Aumentativo e Diminutivo

  Substantivos e adjetivos podem apresentar flexões aumentativas ou diminutivas. O sufixo "ão", por exemplo, normalmente indica um aumentativo, mas existem diversas exceções como "limão" e "mamão" que devem ser consideradas.

- Etapa 5: Redução de Sufixo

  Essa etapa é a que possui o maior número de regras e seu objetivo é retirar terminações de classes gramaticais. Caso faça a remoção, o algoritmo avança diretamente à etapa 8 pulando as etapas 6 e 7.

- Etapa 6: Redução de Sufixo de Verbo

  As flexões verbais do idioma português são muito variadas, o que ocasiona a necessidade de criação de muitas regras para essa etapa. Essas variações se fazem ainda maiores considerando os verbos irregulares e as flexões de tempo, número e modo.

- Etapa 7: Remoção de Vogal

  Essa etapa só será executada caso a palavra não tenha perdido radical nas etapas 5 ou 6. Seu objetivo é remover as vogais "a", "e" e "o" remanescentes das etapas anteriores e deixar apenas a raiz da palavra.

- Etapa 8: Remoção de Acentos

  Algumas palavras que apresentam o mesmo radical podem possuir acentuação na raiz como "catálogo" e "catalogado". Isso criaria dois radicais distintos para o mesmo contexto mofológico. Assim, surge a necessidade dessa etapa de remoção de acentos.

Na figura 2.3 é apresentado um fluxograma que mostra o funcionamento do algoritmo RSLP Stemmer. Segundo Lopes (2004), o RSLP Stemmer apresenta desempenho superior em praticamente todas as classes sintáticas quando comparado ao Porter[1] que também é utilizado para a stemização de palavras. Apesar da qualidade do algoritmo, algumas incertezas podem ocorrer, o que é bastante comum em um idioma complexo como o português.

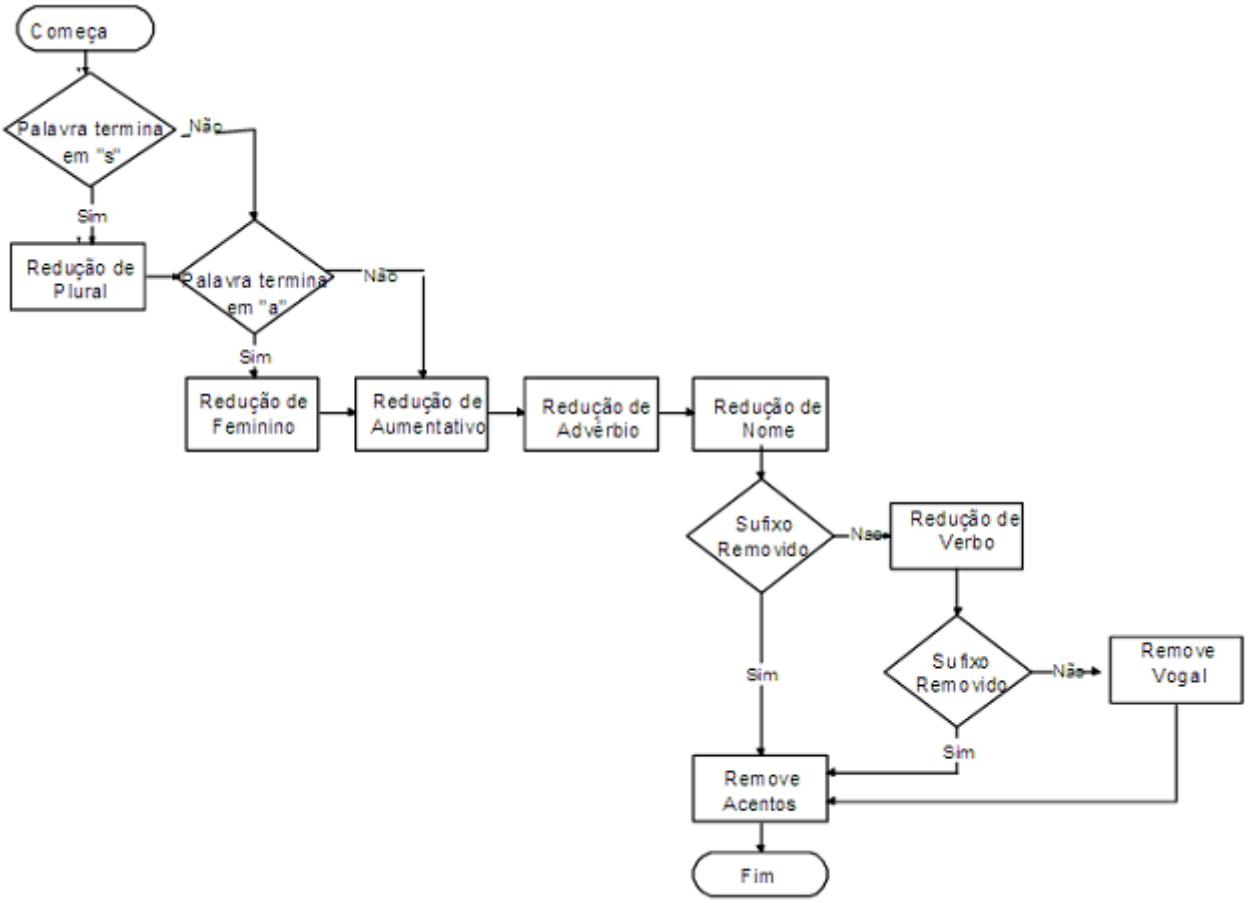

Figura 2.3: Sequência de etapas para o algoritmo de stemização RSLP (Lopes, 2004).

2.7 *Thesaurus*

Thesaurus é um repositório dinâmico e controlado de conhecimentos e temas em forma de lista e organizado de modo que os termos catalogados apresentem relações entre si e nenhum vocábulo listado seja visualizado de maneira isolada. Diversas relações podem

[1]http://tartarus.org/martin/PorterStemmer

ser estabelecidas como hierarquias, equivalências e associações entre os termos indexados. Normalmente documentam contextos restritos e são chamados de microthesaurus, sendo utilizados em coleções de alta especificidade e assuntos bem definidos. Existem também os *thesaurus* utilizados para catalogar coleções genéricas cujo tema pode ser extremamente distinto e a coleção conter uma grande variedade de documentos; esses são os macrothesaurus.

Um uso bastante comum do *Thesaurus* é na produção de documentos, na sugestão de temas, termos e categorizando assuntos. No entanto, é usual sua utilização em sistemas de recuperação da informação a fim de auxiliar a catalogação dos termos indexados. Ao utilizar um *thesaurus* na recuperação da informação, ele ganha um papel de mediador entre a pesquisa do usuário e os termos utilizados na composição dos documentos da coleção.

Segundo Rodrigues (2014),

> O *thesaurus* pode auxiliar o usuário a realizar suas pesquisas, pois quando compreendem a sua estrutura e funcionalidade podem identificar os termos utilizados pelos indexadores em um sistema de recuperação da informação, tanto por digitação, quanto por navegação, no caso deste último, oferecendo autonomia no que se refere ao uso do ambiente digital.

Um *thesaurus* difere de um dicionário já que não traz a definição dos termos, mas apresenta palavras que exprimem ideias semelhantes ou em alguns casos grupos de vocábulos de sentido contrário. No entanto, não deve ser entendido como uma simples listagem de sinônimos mas como um agrupamento de contextos. Podendo ajudar o autor a escolher o vocábulo mais adequado para a produção de um texto ou auxiliando o usuário a localizar uma informação na coleção. Muito popular no ramo da ciência da informação e da biblioteconomia, pode ser utilizado sobre a mesma coleção tanto no âmbito do controle terminológico dos textos redigidos quanto no trabalho do usuário que quer chegar à informação desejada. Vários projetos estão sendo desenvolvidos para o português do Brasil, Entre os mais utilizados destacam-se o TeP 2.0 e a WordNet.Br[2]. A WordNet.Br é um *thesaurus* que possui apenas verbos. Já o TeP 2.0 conta com um maior número de palavras entre verbos, adjetivos, advérbios e substantivos.

[2]http://www.nilc.icmc.usp.br/wordnetbr

2.7.1 TeP 2.0

O TeP 2.0 (Thesaurus para o português do Brasil) é uma base de dados apresentada em um arquivo em formato de texto que contém termos relacionados com seus sinômimos e antônimos. No arquivo, cada linha é um registro que segue o padrão:

Identificador. [Classe gramatical] {Lista de sinônimos} <Identificador de antônimo>

O identificador é um número seguencial que diferencia uma linha da outra. O segundo parâmetro do registro é a classe gramatical. O terceiro parâmetro contém uma lista em que todas as palavras possuem relação de sinônimos entre si, habitualmente chamados de *synsets*. Já o quarto parâmetro é um identificador que relaciona o registro com outro grupo de ideias antônimas, mas não é encontrado em todas as linhas.

O TeP 2.0 é fundamental na análise textual e no estabelecimento de relações entre palavras. Segundo Maziero; Pardo (2008), ele é de vital importância para aplicativos de Processamento de Línguas Naturais, pois consiste em um primeiro passo para se lidar automaticamente com as palavras e seus sentidos.

A base do TeP 2.0 pode ser consultada pela internet[3]. A figura 2.4 apresenta a interface da plataforma e suas principais atribuições.

Figura 2.4: Interface *on line* de consulta do TeP 2.0 (Maziero; Pardo, 2008).

O TeP em sua versão 2.0 possui 19.888 registros de grupos de sinônimos e 18.163 registros de antônimos, totalizando 44.678 palavras. Sua construção é manual, o que

[3]http://www.nilc.icmc.usp.br/tep2

desencadeia diversas possibilidades de erros e incompatibilidades de dados. Assim, se faz necessário um processo contínuo de correções e evolução de sua base.

2.8 Métricas de Recuperação de Informação

Sem que se faça uma avaliação do sistema de recuperação da informação não é possível saber se ele cumpre seu objetivo de fornecer ao usuário a informação desejada. Dentro de uma mesma coleção podem ser aplicados vários métodos de retornos de consulta e para saber qual o mais adequado usa-se uma métrica para avaliar a qualidade dos resultados quanto aos documentos retornados, verifica-se se são ou não relevantes para a consulta.

Outro ponto que deve ser observado é a quantidade de documentos devolvidos pela busca que não são relevantes ao usuário. Esse número deve ser minimizado ao máximo já que é um indicação negativa ao funcionamento do sistema de recuperação da informação. Além disso, alguns documentos, mesmo sendo relevantes, não são recuperados devido à variação linguística destoar da pesquisa feita pelo usuário.

Assim, a avaliação de um sistema de recuperação da informação deve ser pautada na relação quantitativa entre os documentos devolvidos pelo sistema e os documentos desejados pelo usuário que realizou a pesquisa.

Normalmente, para realizar o cálculo da métrica estabelecida sobre cada consulta, é realizada uma comparação do conjunto retornado e o conjunto almejado sugerido por pessoas que contribuam para a avaliação. Para Júnior; Tarapanoff (2006), os sistemas de recuperação de informação, além de buscarem atender às demandas informacionais dos usuários, dependem destes para que a qualidade dos seus serviços seja reconhecida. Isso pode ser encarado como um problema já que cada usuário tem seu próprio critério de avaliação, mas médias podem ser definidas em um grupo e assim qualificar um sistema.

A revocação (*recall*) é a métrica utilizada para quantificar o desempenho do sistema de recuperação da informação em relação ao seu potencial de recuperar documentos úteis ao usuário segundo sua consulta. A equação 2.7 apresenta a forma que esse índice deve ser calculado. Quanto mais o resultado se aproxime de 1, melhor é o desempenho do sistema já que tende a recuperar a grande maioria dos itens relevantes à pesquisa realizada.

$$revocação = \frac{N^o\ total\ de\ documentos\ relevantes\ recuperados}{N^o\ total\ de\ documentos\ relevantes\ da\ coleção} \tag{2.7}$$

A precisão é a métrica utilizada para quantificar o desempenho do sistema de recuperação da informação em relação à sua eficiência em reprimir o retorno de documentos irrelevantes à consulta. A equação 2.8 apresenta a forma com que é calculado esse índice. Quanto mais esse valor se aproxima de 1, melhor é a taxa de precisão. O que significa que a probabilidade de um documento retornado ser relevante é maior e melhor é o desempenho do sistema nesse quesito de recuperação de itens pertinentes.

$$precisão = \frac{N^o\ total\ de\ documentos\ relevantes\ recuperados}{N^o\ total\ de\ documentos\ recuperados} \tag{2.8}$$

3 Projeto

3.1 Introdução

O objetivo da execução do projeto é construir um sistema de busca para as atas do CONSU da UFJF e verificar o funcionamento do TeP 2.0 perante essa coleção de documentos. Para tal, o sistema foi desenvolvido em PHP[4], uma linguagem livre e muito utilizada em aplicações dinâmicas na *World Wide Web*. Para controlar o banco de dados, foi utilizado o MySQL[5] que é um sistema de gerenciamento de banco de dados que utiliza a linguagem SQL (*Structured Query Language*) para comunicação. Ele foi escolhido por ser de fácil utilização e um dos mais populares do mundo. Além disso, foram utilizados diversos processos comuns aos sistemas de recuperação da informação com stemização e utilização de *thesaurus* que serão discutidos a seguir.

3.2 Interface da Ferramenta

O objetivo da interface é ser o mais simples possível. Assim, esse projeto apresenta apenas três páginas: a página de gerenciamento de atas, a página de busca que utiliza o TeP 2.0 e a página de busca que não o utiliza. Para acessar essas páginas, o sistema possui um menu de opções conforme a figura 3.1.

Figura 3.1: Menu de opções do projeto

A página de gerenciamento de atas permite ao usuário administrador do sistema realizar o cadastro de novas atas e a visuzalização das atas cadastradas. Em caso de erros de inclusão, é possível fazer a exclusão de registros indevidos. A figura 3.2 apresenta a tela de gerenciamento de atas do sistema.

[4]http://php.net
[5]https://www.mysql.com

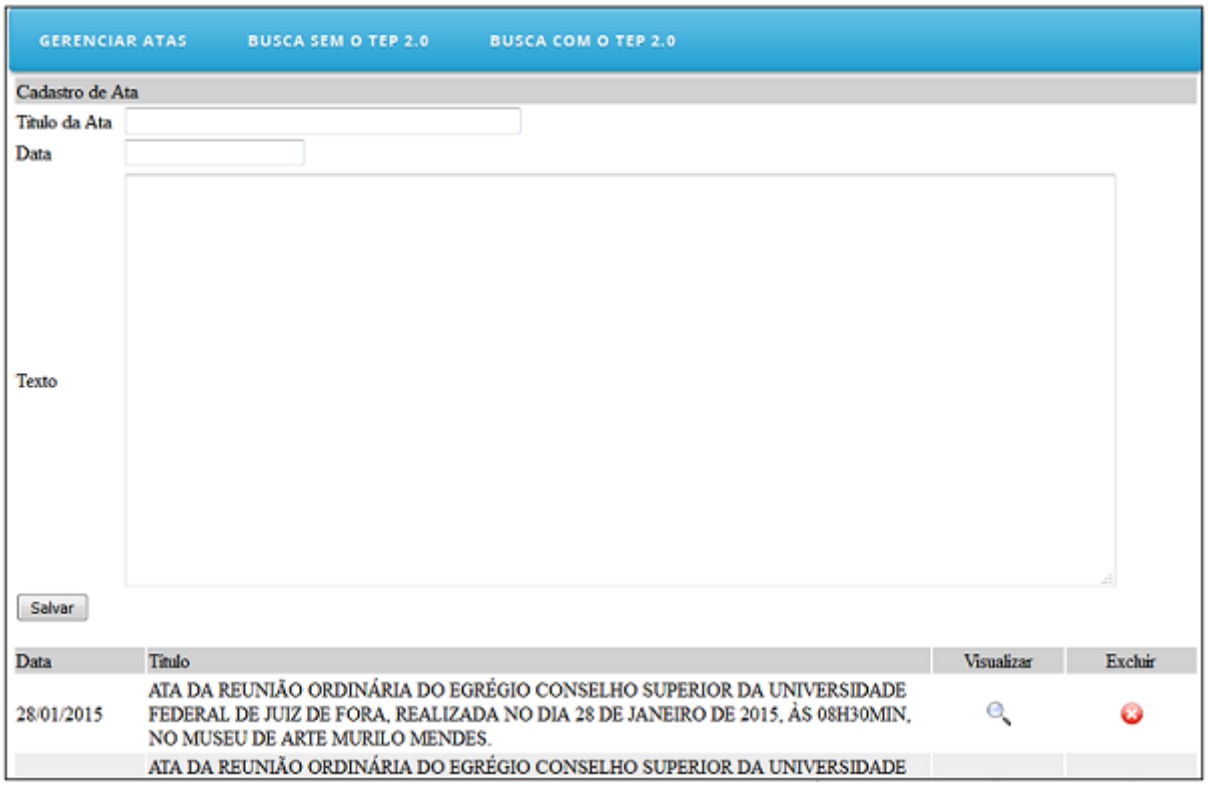

Figura 3.2: Tela de gerenciamento de atas

A tela de busca que não utiliza o TeP 2.0 apresenta apenas um campo para digitar a pesquisa e um botão de submissão, conforme a figura 3.3.

Figura 3.3: Tela de busca de atas que não utiliza o TeP 2.0

A tela de busca que utiliza o TeP 2.0 é bastante semelhante à tela que não o utiliza. Para melhor diferenciação, foi incluído um título em cinza, o que aprimora a utilização para o usuário, como pode ser observado na figura 3.4.

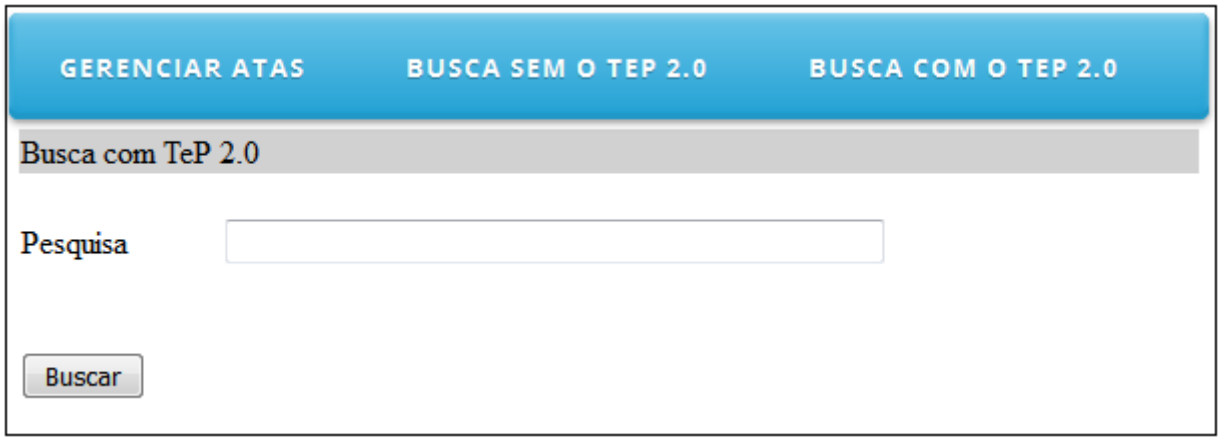

Figura 3.4: Tela de busca de atas que utiliza o TeP 2.0

3.3 Modelagem da Ferramenta

A figura 3.5 apresenta o diagrama de classes do projeto que foi criado utilizando o programa Astah[6]. A seguir, na tabela 3.1, são apresentadas todas as tabelas do banco de dados juntamente com a descrição das atribulições de cada campo armazenado.

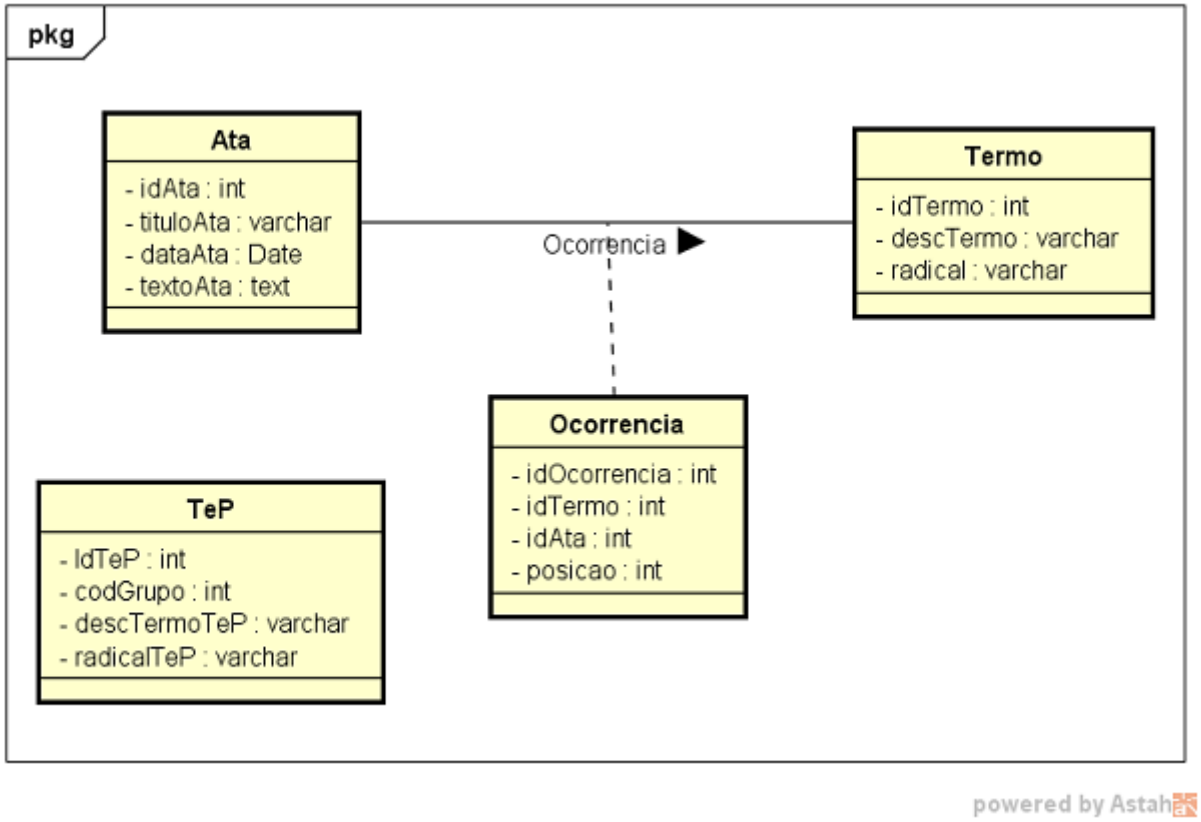

Figura 3.5: Diagrama de classes do projeto

[6]http://astah.net

Tabela 3.1: Tabelas do diagrama de classes

Ata	idAta	Identificador de cada registro da tabela.
	tituloAta	Título da ata descrito pelo redator.
	dataAta	Data em que ocorreu a reunião.
	textoAta	Descreve tudo o que aconteceu na reunião.
Termo	idTermo	Identificador de cada registro da tabela.
	descTermo	Palavra contida no documento que foi indexada.
	radical	Palavra que passou pelo processo de stemização.
Ocorrencia	idOcorrencia	Identificador de cada registro da tabela.
	idTermo	Faz referência a um registro na tabela "Termo".
	idAta	Faz referência a um registro na tabela "Ata".
	posicao	Armazena a posição de ocorrência no documento.
TeP	idTeP	Identificador de cada registro da tabela.
	codGrupo	Grupo de sinônimos propostos pelo TeP 2.0.
	descTermo	Palavra catalogada no TeP 2.0.
	radicalTeP	Radical da palavra do TeP 2.0 stemizada pelo RSLP.

3.4 Processo de Indexação Utilizado na Ferramenta

Após a inserção da ata no banco de dados, inicia-se o processo de indexação das palavras do documento. Para tal, a função de indexação recebe uma lista contendo todas as palavras da ata. Percorrendo essa lista, para cada termo analisado são realizados os seguintes procedimentos:

1. Verifica-se a catalogação do termo, ou seja, se já foi cadastrado na tabela "Termo". Essa tabela armazena os distintos termos da coleção.

2. Caso a palavra não tenha sido catalogada é feito o cadastro do termo.

3. Faz-se o registro de ocorrência da palavra na tabela "Ocorrência" armazenando a posição em que o termo foi encontrado no documento.

3.5 Algoritmo de Stemização da Ferramenta

Antes de iniciar o processo de stemização, foi necessária uma transformação da palavra para que todas utilizem caracteres minúsculos, já que em caso de comparações com as exceções listadas poderiam ocorrer incompatibilidades pois a linguagem PHP considera diferente o mesmo termo escrito com letras maiúculas e minúsculas. O algoritmo a seguir é uma implementação do algoritmo RSLP. Originalmente, esse algoritmo foi escrito em

C. Esta é uma tradução própria desenvolvida em PHP.

```
1  public function gerarStem($palavra){
2      $palavra = strtolower($palavra);
3      $palavra = $this->regraPlural($palavra);
4      $palavra = $this->regraFeminino($palavra);
5      $palavra = $this->regraAumentativo($palavra);
6      $palavra = $this->regraAdverbio($palavra);
7      $palavra1 = $this->regraNome($palavra);
8      //verifica qual palavra deve ser usada
9      if($palavra1 == $palavra){
10         $palavra1 = $this->regraVerbo($palavra);
11         if($palavra1 == $palavra){
12             $palavra1 = $this->regraVogal($palavra);
13         }
14     }
15     //verifica qual palavra deve ser usada
16     if($palavra1 == $palavra){
17         $palavra = $this->retirarAcentos($palavra);
18     }else{
19         $palavra = $this->retirarAcentos($palavra1);
20     }
21     return $palavra;
22 }
```

3.6 Armazenamento do TeP 2.0

Para fazer o armazenamento, foi criado um *script* fora do contexto do sistema. Uma vez incluído o TeP 2.0, esse *script* não deve ser mais utilizado. A base do TeP 2.0 é disponibilizada em arquivo de texto onde cada linha é um registro de grupo de sinônimos com várias palavras. Assim, foi criado um ponteiro para percorrer cada linha do arquivo e para cada linha são realizados os seguintes procedimentos:

1. Eliminação da pontuação que separa as palavras.
2. Eliminação da partícula "-se" que acompanha diversas palavras do TeP 2.0.
3. Eliminação dos marcadores de classificação gramatical dos grupos de sinônimos ([Verbo], [Adjetivo], [Advérbio], [Substantivo]).
4. Eliminação dos marcadores de antônimos.
5. Alocação das palavras em uma lista.

6. Para cada palavra da lista, realizar os procedimentos:

    - Gerar o radical utilizando o RSLP.
    - Designar valores das variáveis: codGrupo, descTermoTeP e radicalTep.
    - Inserir registro na tabela TeP.

3.7 Buscas Utilizando a Ferramenta

O processo de busca no sistema é bastante simples. O usuário digita no campo de pesquisa o conjunto de palavras-chave que deseja localizar na coleção de atas. Ao submeter o formulário, são executados os seguintes procedimentos:

1. Criação de uma lista com as palavras pesquisadas.
2. Stemização das palavras pesquisadas.
3. Pesquisa na tabela "Ocorrencia" os radicais de termos presentes nas atas.
4. Criação de uma lista de atas que possuam os radicais referentes as palavras buscadas.
5. Cálculo do índice TF-IDF para cada ata da lista de atas retornadas.
6. Ordenação das atas de acordo com o índice TF-IDF.
7. Apresentação do resultado da pesquisa ao usuário.

A sequência apresentada anteriormente refere-se à busca que não utiliza o TeP 2.0. A única diferença para a busca que o utiliza é que na fase inicial do processamento são incluídas palavras cujos radicais são apresentados como sinônimos dos radicais das palavras pesquisadas pelo usuário.

4 Resultados

Tabela 4.1: Amostra de resultados de testes no projeto

Nº da Pesquisa	1	2	3	4	5
Nº de atas relevantes da coleção	10	10	7	14	10
Nº de atas retornadas com TeP	14	13	5	25	10
Nº de atas retornadas sem TeP	7	7	5	1	8
Nº Relevantes com TeP	9	9	5	14	8
Nº Relevantes sem TeP	6	6	5	1	8
Revocação com TeP	0.9	0.9	0.71	1	0.8
Revocação sem TeP	0.6	0.6	0.71	0.07	0.8
Precisão com TeP	0.64	0.69	1	0.56	0.8
Precisão sem TeP	0.85	0.85	1	1	1

A tabela 4.1 e as figuras 4.1 e 4.2 apresentam os resultados obtidos no projeto. Através desses resultados e de todos os outros dados levantados pode-se observar que a utilização do TeP 2.0 trouxe vantagem significativa em várias pesquisas no quesito revocação de documentos, uma vez que proporcionou aumento do número de atas relevantes retornadas. No entanto, isso não ocorreu em todas as buscas, mas pelo menos foi minimamente igual aos resultados alcançados nas buscas que não utilizaram o TeP 2.0. Já no quesito precisão, o projeto se comportou de maneira inversa: a utilização do TeP 2.0 apresentou uma piora nos resultados uma vez que aumentou tanto o retorno de documentos relevantes quanto os que não são de interesse do usuário. No entanto, não foi possível avaliar a variação desses dois indicadores já que os resultados encontrados apresentaram uma inconstância, visto que o número de experimentos foi pequeno e o tamanho da base de atas ainda é limitado, contando com apenas 25 atas referentes aos anos de 2015 e 2016. Portanto, é preciso pensar tanto na ampliação da base de atas como do número de experimentos.

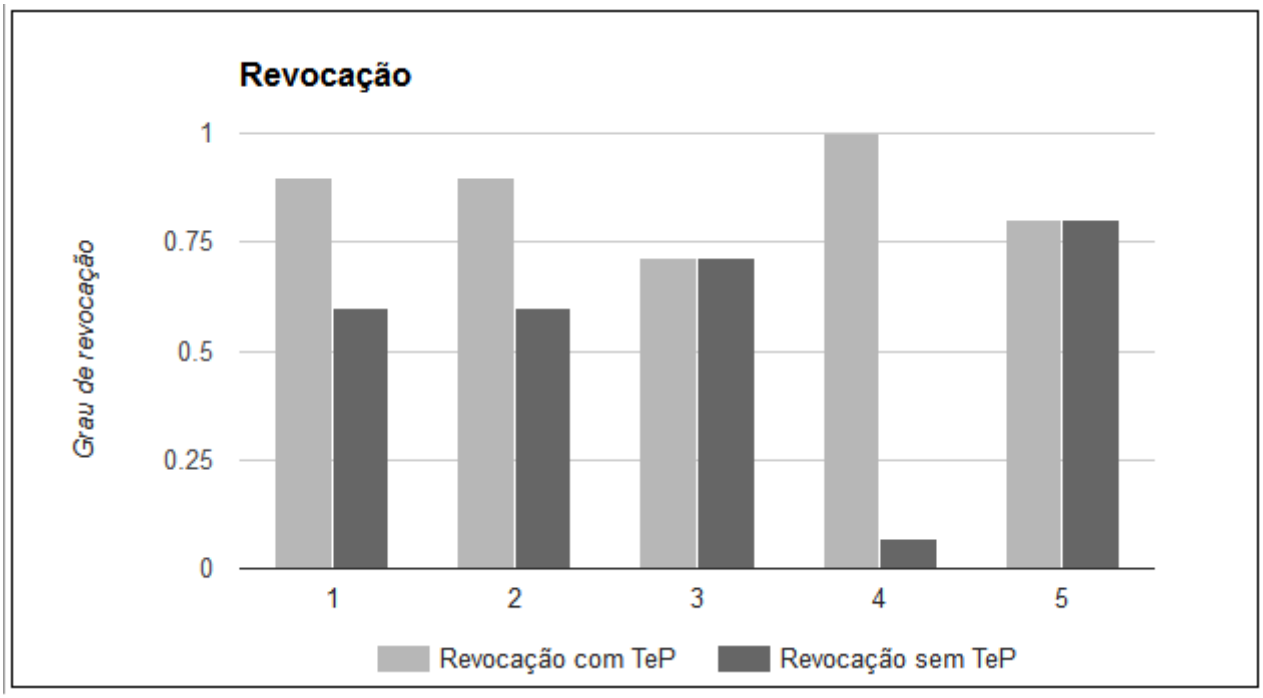

Figura 4.1: Revocação do projeto

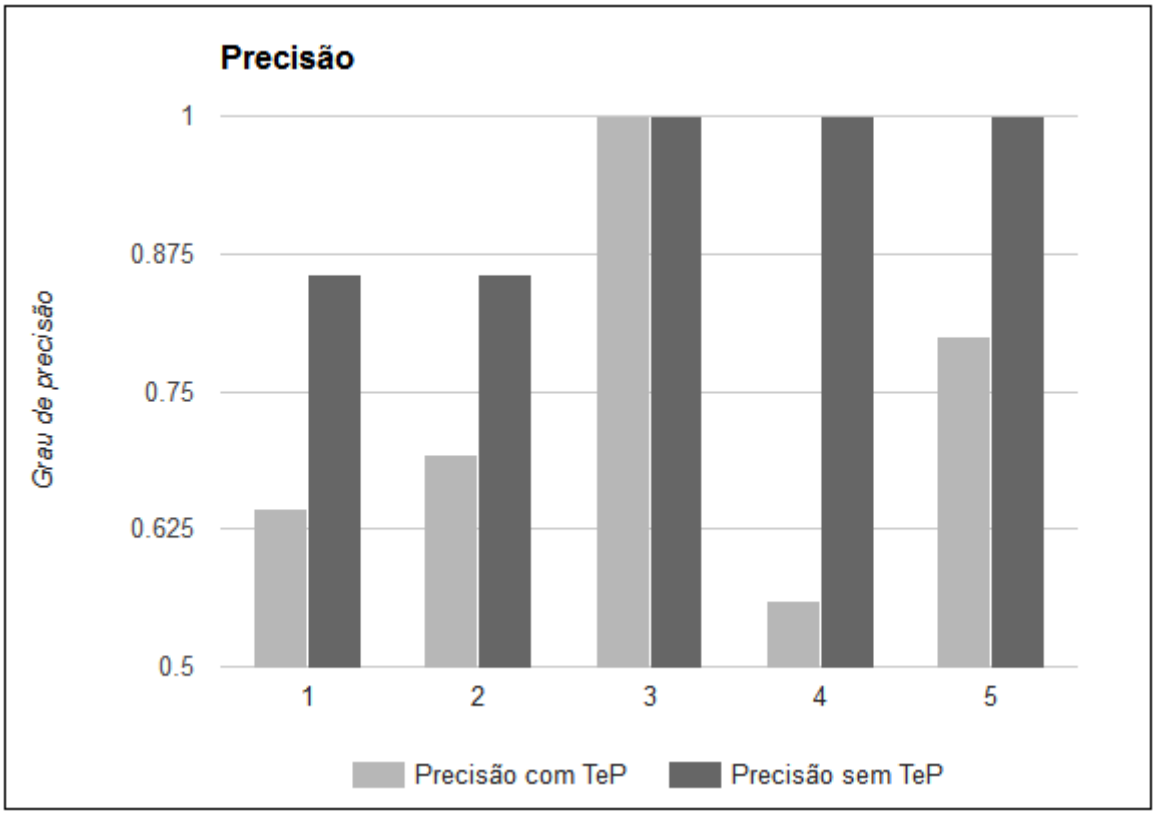

Figura 4.2: Precisão do projeto

5 Conclusão

Outros modelos semânticos, como ontologias e dados ligados, podem ser utilizados em sistemas de RI. Uma ontologia é um modelo de dados semântico que representa um conjunto de conhecimentos dentro de um contexto e as relações que podem ser estabelecidas entre os dados armazenados. Dados ligados é um conjunto de normas utilizadas para relacionar dados na web. Apesar de ambos poderem ter sido utilizados neste trabalho o uso de *thesaurus* se mostrou mais adequado já que o maior objetivo era estabelecer apenas relação de sinonímia entre os termos armazenados.

Em relação ao desempenho de uso da ferramenta, notou-se uma lentidão de resposta, em especial na pesquisa que utilizada o Tep 2.0. Isso se deve ao número excessivo de processamento que se faz necessário no processo de busca de radicais na base do Tep 2.0 e sua localização na base de indexação. É interessante que para próximos trabalhos seja realizada uma refatoração dos processos utilizados no projeto. A base do TeP 2.0 também necessita de constante atualização já que foi desenvolvida manualmente, o que pode ocasionar erros e incertezas.

Para melhorar o funcionamento geral do sistema, pode ser realizada uma atualização dos arquivos de regras de stemização de acordo com as palavras contidas nos documentos tendo em vista que a grande maioria das palavras ocorre diversas vezes na coleção. Em muitas das buscas ocorreram incertezas quanto aos radicais gerados, o que é comum em processos de stemização. Os processos de indexação também devem ser aprimorados já que nesse projeto houve uma catalogação total das palavras dos documentos.

Assim, constatou-se que o uso do TeP 2.0 pode proporcionar benefícios em sistemas de recuperação de informação quando bem empregado e avaliado diante da coleção desejada. Porém, sua grande desvantagem é não ser um serviço web o que obriga o desenvolvedor a fazer atualizações manuais de sua base.

Por hora, a ferramenta não está disponibilizada na web, apesar de estar preparada pra isso.

Bibliografia

Arantes, A. G. Q. **Implementação do módulo de indexação e consulta para ser agregado ao metabuscador do portal do ceulp/ulbra**. Palmas, 2005.

Baeza-Yates, R.; Ribeiro-Neto, B. **Recuperação da informação: Conceito e tecnologia das máquinas de busca (2ª edição)**. Porto Alegre, 2011. Bookman Editora LTDA.

Corrales, J. D. **Ayudante técnico de informática de la junta de andalucía(2ª edição)**. Belo Horizonte, 2005. Editorial MAD, S.L.

Dias, M. A. L.; de Gomensoro Malheiros, M. **Extração automática de palavras-chave de textos da língua portuguesa**. Lajeado, [2006?]. Disponível em: <http://wsl.softwarelivre.org/2005/0020/20.pdf>. Acesso em: 12 jun. 2016.

Ferneda, E. **Recuperação da informação: Análise sobre a contribuição da ciência da computação para a ciência da informação**. São Paulo, 2002.

de Araújo Júnior, R. H.; Tarapanoff, K. **Precisão no processo de busca e recuperação da informação: uso da mineração de textos**. Brasília, 2006.

Longhi, M. T.; Behar, P. A.; Bercht, M. ; Simonato, G. **Investigando a subjetividade afetiva na comunicação assíncrona de ambientes virtuais de aprendizagem**. Porto Alegre, 2009. XX Simpósio Brasileiro de Informática na Educação.

Lopes, M. C. S. **Mineração de dados textuais utilizando tácnicas de clustering para o idoma português**. Rio de Janeiro, 2004.

Maziero, E. G.; Pardo, T. A. S. **A interface de acesso ao tep 2.0 - thesaurus para o português do brasil**. São Carlos, 2008.

Neubert, M. S. **Algoritmos distribuídos para a construção de arquivos invertidos**. Belo Horizonte, 2000.

Oliveira, E. L. S. **Uma investigação de aspectos da classificação de tópicos para textos curtos**. João Pessoa, 2015.

Orengo, V. M.; Huyck, C. **A stemming algorithm for the portuguese language**. In: Proceedings of the SPIRE Conference, Laguna de San Raphael, Chile, 2001.

Rodrigues, A. M. **O uso do tesauro na arquitetura da informação em websites**. Brasília, 2014.

Tsuji, G. K.; Kamaura, L. T. **Integrando recuperação de informação em banco de dados com hibernate search**. São Paulo, 2008.

Wikimedia. **Stemização**. Disponível em: <http://pt.wikipedia.org/wiki/Stemização>. Acesso em: 20 mar. 2016.