UNIVERSIDADE DO SUL DE SANTA CATARINA

EDUARDO MATEUS ALBERTON

THIAGO DE SOUZA

SISTEMA PARA EXTRAÇÃO DE INDICADORES A PARTIR DA LISTA DE REFERÊNCIAS DE TRABALHOS CIENTÍFICOS

FLORIANÓPOLIS

2015

EDUARDO MATEUS ALBERTON
THIAGO DE SOUZA

SISTEMA PARA EXTRAÇÃO DE INDICADORES A PARTIR DA LISTA DE REFERÊNCIAS DE TRABALHOS CIENTÍFICOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas de Informação da Universidade do Sul de Santa Catarina, como requisito parcial à obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Flávio Ceci, Dr.

FLORIANÓPOLIS
2015

EDUARDO MATEUS ALBERTON
THIAGO DE SOUZA

SISTEMA PARA EXTRAÇÃO DE INDICADORES A PARTIR DA LISTA DE REFERÊNCIAS DE TRABALHOS CIENTÍFICOS

Este Trabalho de Conclusão de Curso foi julgado adequado à obtenção do título de Bacharel em Sistemas de Informação e aprovado em sua forma final pelo Curso de Graduação em Sistemas de Informação da Universidade do Sul de Santa Catarina.

Florianópolis, 18 de Novembro de 2015.

Professor e orientador Flávio Ceci, Dr.
Universidade do Sul de Santa Catarina

Prof. Aran Bey T. Morales, Dr.
Universidade do Sul de Santa Catarina

Prof. Mauricio Botelho, M.Eng.
Universidade do Sul de Santa Catarina

RESUMO

Extrair informações em bases não estruturadas é um problema conhecido a muito tempo. Dados armazenados em documentos textuais não são recuperados facilmente como informações em um banco de dados por exemplo. Considerando este como o problema pressuposto, este trabalho tem como objetivo desenvolver um sistema de extração de informação juntamente com a geração de indicadores voltado à área acadêmica, realizando a extração de dados de referências biliográficas elaboradas no formato exigido pela ABNT. Isso só é possivel de ser realizado, por possuir uma regra estipulada, onde os elementos devem respeitar posições e delimitações pontuais. O uso de expressões regulares para encontrar entidades declaradas em uma determinada sentença tambem ajuda na identidicação dos elementos em arquivos textuais, possibilitando assim, a extração da informação desejada. O sistema desenvolvido para solução proposta realiza a leitura e extração de autores e ano de publicação das referências bibliográficas adicionadas ao sistema. Os documentos inseridos no sistema são trabalhos de conclusão de curso da Universidade de Sul de Santa Catarina, dos cursos de Ciências da Computação e Sistemas de Informação além de artigos científicos. O principal objetivo da extração dos dados citados é apresentar esta informação de forma sintética e pura, auxiliando na visualização da credibilidade dos documentos e do período no qual foi buscado seu conteúdo base. Desta forma, o sistema captura os dados de arquivos textuais, e os apresenta de forma agrupada através da geração de indicadores. O sistema foi avaliado obtendo um índice de acerto de 90,92 % na extração de autores, e de 97,75 % para a extração do ano de publicação, de um total de 535 referências.

Palavras-chave: Extração de Informação, Indicadores de Publicações Acadêmicas, Referências Bibliográficas, Expressões regulares.

Sumário

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

1 INTRODUÇÃO

Papa Filho e outros (2002) apontam que cada vez mais pessoas estão em busca de informações para auxiliar na tomada de decisão. Os sistemas de informação têm como objetivo facilitar e agilizar as atividades humanas, dinamizar a capacidade de tomar decisões e refinar estratégias de relacionamento.

Os sistemas de informação com base no apoio à tomada de decisões, como um conjunto de componentes relacionados, podem afetar diretamente na estratégia de uma organização. A informação ocupa um espaço importantíssimo na formulação de estratégias de mercado como também no acompanhamento das operações empresariais. (PAPA FILHO e outros, 2002).

Da Silva e outros (2003) afirmam através de indicadores gerados por sistemas de apoio a decisão, a partir de uma massa de dados, é possível obter medidas, e com isso, decidir se algo esta sendo realizado da forma esperada ou não.

> Indicadores resultam da agregação de dados brutos e processados, mas podem ser agregados novamente para formar índices complexos. Índices são medidas com alto nível de agregação, que combinam os indicadores mais importantes para descrever o desempenho de uma instituição, região ou setor econômico. Os índices normalmente simplificam e traduzem sistemas complexos através de um número único, que pode ser útil à tomada de decisão (DA SILVA ; AGOPYAN, 2003, p. 19).

Com a grande disseminação das publicações científicas nas diversas áreas do conhecimento, tornou-se necessário a criação de critérios mais exigentes para avaliação do que é publicado, há uma preocupação dos profissionais que se interessam pela qualidade da informação científica (STREHL, 2005).

> Este modo de avaliação de qualidade feita a partir do impacto das publicações na comunidade científica é denominado no ramo da bibliometria e da cientometria como análise de citações, ou estudo de citações, e tem se difundido mundialmente no âmbito das agências de fomento de pesquisa (STREHL, 2005, p. 19).

Dentre o ramo de indicadores gerados a partir de publicações cientificas, existem alguns sistemas de qualificação e estratificação já consolidados, como o JRC que “ajuda a medir a influência e impacto nos níveis de revistas e de categoria, mostrando a relação entre citações e citados em revistas, apresentando dados quantitativos que suportam uma revisão

sistemática e objetiva dos principais periódicos do mundo, usando uma combinação de impacto e influência“. (THOMSON REUTERS, 2015).

A Coordenação de Aperfeiçoamento de Pessoa de Nível Superior (CAPES) utiliza um conjunto de procedimentos, conhecidos como Qualis para medir a qualidade de produção intelectual dos programas de pós-graduação. Sendo este outro indicador já consolidado. Capes (2014) informa que este indicador foi concretizado com o intuito de atender as necessidades do sistema de avaliação, sendo este baseado em dados disponibilizados por um aplicativo conhecido como Coleta de Dados. (COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR, 2014).

Indicadores de publicações são gerados através de alguma fonte de dados. Bordin e outros (2013) abordam que uma grande parte destas fontes de dados atual está presente em fontes não estruturadas, sendo esse o motivo que levou ao surgimento da área de descoberta de conhecimento em texto.

Segundo Ebecken, Lopes e Costa (2005, apud BORDIN e outros 2013), descoberta de conhecimento em texto é “um conjunto de técnicas e métodos que têm como função explicitar informações e conhecimento que estão implícitos em textos”.

O objetivo deste trabalho é utilizar a extração de informação juntamente com a geração de indicadores, aplicada em um ambiente científico, utilizando referências bibliográficas como fonte de extração de dados.

> Ao se trabalhar com dados em bases eletrônicas, pode-se distinguir formas em que é possível representá-las. As páginas da Internet são consideradas dados “semiestruturados”, de caráter intermediário, ou seja, que apresentam “alguma estrutura”. Têm-se ainda os dados “estruturados”, como, por exemplo, aqueles presentes nos bancos de dados relacionais, e os “não-estruturados”, como, por exemplo, o texto livre. (ALMEIDA, 2002, p.4).

Como as referências bibliográficas são informações semiestruturadas, ou seja, possuem algum tipo de estrutura, possibilitam, desta forma, a definição dessas regras para que as informações nelas contidas sejam extraídas da forma correta.

1.1 PROBLEMA DE PESQUISA

Para realizar um trabalho científico, é necessário obter muita informação já obtida por outros autores. Prodanov e outros (2013) informam que é necessário utilizar um referencial bibliográfico bem fundamentado com o objetivo de contextualizar a pesquisa e dar um embasamento para o que está sendo apresentado. Extrair os dados referenciais de um único documento, já é bastante trabalhoso. Realizar essa extração em uma grande escala, então, avaliando a qualidade de informação, pesquisando por suas fontes e fundamentos e histórico de publicação ou citação, torna-se uma tarefa muito difícil para um humano.

Lopes (2004) explica a importância de se ter um bom referencial teórico e de uma boa metodologia para desenvolvimento de um trabalho. O autor explica que, através do referencial teórico, a atividade científica não só organiza e objetiva o seu contexto, como também ajuda a alcançar os objetivos da pesquisa.

Além de realizar a extração destes dados, precisa-se armazená-los em algum lugar, para possibilitar a realização do processamento e a disseminação dos mesmos.

É difícil o desenvolvimento de aplicações que necessitem capturar e manipular informações diretamente do conteúdo digital, como, por exemplo, na busca de conteúdo em vídeos, sem utilizar seus metadados descritivos – que são trechos de informação textual associados aos vídeos, geralmente compreensível apenas por seres humanos. (BATISTA; SCHWABE, 2009).

Se equiparar isso para um universo manual, torna-se uma tarefa extremamente difícil de ser realizada por pessoas, sendo até possível de ser feita, entretanto, em uma escala muito menor e de forma muito mais trabalhosa.

Além de extrair, processar e armazenar esses dados, é preciso também, apresentar esses dados utilizando indicadores, para facilitar a visualização e compreensão dos dados.

Da Silva e outros (2003) afirmam, através de indicadores gerados por sistemas de apoio à decisão, a partir de uma massa de dados, é possível obter medidas.

Entretanto, como se pode apresentar a informação sobre os dados, para que tenha utilidade e gere interesse de conhecimento e para o usuário?

1.2 OBJETIVOS

Esta seção é reservada a apresentar o objetivo geral deste trabalho e seus objetivos específicos.

1.2.1 Objetivo geral

Este trabalho tem como objetivo desenvolver uma aplicação para extrair indicadores a partir de uma lista de referências de trabalhos científicos.

1.2.2 Objetivos específicos

Como objetivos específicos, o presente trabalho visa:

- Identificar padrões de referências a partir da ABNT;

- desenvolver um módulo para reconhecimento de padrões a partir da lista de referências;

- propor um modelo de dados que suporte o armazenamento das informações extraídas;

- desenvolver um sistema web que implemente e que represente a solução para o problema descrito;

- avaliar o módulo desenvolvido a partir da sua taxa de acerto.

1.3 JUSTIFICATIVA

Segundo Wives e Loh (1998), muitas das informações disponíveis para acesso rápido e fácil não estão em formatos que possam ser tratados por meios computacionais (imagens, textos, vídeos, gráficos).

Para Tan (1999, apud WIVES 2004), mais de 80% das informações encontradas, atualmente, encontram-se em formato textual. Assim, é importante que os métodos de análise e processamento de dados sejam direcionados para esse tipo de informação, utilizando documentos.

Utilizando expressões regulares, é possível criar modelos para extrair a informação de fonte de dados, esses modelos são chamados de padrões que visam a realizar uma normalização das palavras existentes nesses documentos. Para Wives (2004), essas técnicas de pré-processamento, com fins de normalização, são provenientes de áreas como "Processamento de Linguagem Natural" e "Sistemas de Recuperação de Informações".

Loh (2001) afirma que esses modelos são estruturas capazes de representar objetos e ideias através de regras de extração, que levam em conta problemas de vocabulário derivado, permitindo modelar melhor o conteúdo de documentos. Dessa forma, um conjunto de regras constitui a forma de extração.

Esse enorme número de informações armazenadas ao longo dos anos foram gravadas com o intuito de serem recuperadas e terem uma utilidade. Essa análise de dados pode ser feita por meio de indicadores que buscam aplicar medidas sobre os dados tratados, de forma a auxiliar o ser humano nas suas tomadas de decisões, tendo uma visão mais ampla do conteúdo relacionado.

A geração de indicadores voltado ao contexto das referências bibliográficas servem para, além de apresentar de forma sucinta os as informações contidas, saber se um arquivo científico (artigo, tese, etc.) encontra-se no estado da arte, ou seja, se o material utilizado para desenvolvimento do arquivo é o mais indicado para o tema e, também ,para ter conhecimento da atualidade que encontram-se as referências de determinado trabalho.

Esta pesquisa visa a abordar o processo de descoberta do conhecimento em dados em textos, com o principal objetivo de revelar relacionamentos interessantes que possam ser utilizados para a tomada de decisão na área de referências bibliográficas em trabalhos científicos, aplicando técnicas de agrupamento de informações e geração de indicadores.

1.4 ESTRUTURA DO TRABALHO

O presente trabalho está organizado em seis capítulos, o primeiro e presente capítulo está organizado da seguinte forma: a introdução apresenta como está organizado o cenário de aplicação, seguindo de uma seção de problemática, que procura apresentar os principais problemas e desafios relacionados com o tema de pesquisa, na sequência são apresentados os objetivos gerais e específicos. Na próxima seção, são demonstradas justificativas que procuram ilustrar a importância do tema. Por fim, é construída uma seção sobre a estrutura em que este trabalho se encontra organizado.

O capítulo dois tem como função apresentar um referencial teórico, a fim de suportar os temas relacionados com o objetivo geral deste trabalho.

No capítulo três é apresentado o método de pesquisa que se utiliza como base para o desenvolvimento da pesquisa. No capítulo quatro são apresentados os artefatos de modelagem do protótipo de solução, bem como a metodologia utilizada como base. O capitulo cinco apresenta mais detalhes sobre o sistema desenvolvido bem como a sua avaliação e resultados na aplicação do mesmo.

Por fim, no capítulo seis são apresentadas as conclusões e os trabalhos futuros.

2 REFEÊNCIAL TEÓRICO

Neste capítulo, é apresentado o referencial teórico ao tema deste documento, onde são abordados os temas de extração de informação, contendo os tipos de dados existentes, além de técnicas para extração de informação.

São apresentadas, também, informações referentes a expressões regulares, formatos de referências bibliográficas, além de informações referentes aos indicadores de publicações acadêmicas.

2.1 EXTRAÇÃO DE INFORMAÇÃO

Lopes (2004) aponta que, com o decorrer dos tempos, informações são armazenadas constantemente em textos e base de dados com uma quantidade grande de dados tornou-se difícil consultar informações em textos, levando ao surgimento de técnicas para possibilitar a extração de informação.

A partir do advento do *text mining*, foi possível a extração de informações em textos e, com isso, um novo arsenal de dados pode ser analisado e estudado. A partir dessa descoberta, um grande volume de dados eletrônicos começou a ser gerado e, para isso, foram criadas cada vez mais técnicas de extração automáticas para valorizar e aproveitar este grande número de informações. (LOPES, 2004).

Bordin e outros (2013) informam que existe uma grande quantidade de conhecimento das organizações, armazenadas em forma de dados não estruturados, ou seja, texto bruto. Sendo esse o motivo que fez com que a tecnologia da informação se voltasse para a área de Descoberta de Conhecimento em Texto (*Knowledge Discovery in Text - KDT*).

Segundo Ebecken e outros (2005, apud BORDIN e outros 2013), "Descoberta de conhecimento em texto é um conjunto de técnicas e métodos que têm como objetivo explicitar informações e conhecimento que estão implícitos em textos".

Silva Filho (2009) informa que, através do processo de descoberta de conhecimento em texto (KDT), foi possível administrar essa grande quantidade de dados, que se encontra em forma não estruturada, em forma de conhecimento, algo que, muitas vezes, é inovador para as empresas.

Com isso, pouco a pouco, esses métodos foram sendo adotados pelo pessoal da área de Sistemas de Informação, que descobriu que suas coleções de dados poderiam ser fontes valiosas de conhecimento.

Nesse mesmo contexto, Rezende, Marcacini e Moura (2011, p 1) informam que

> A Mineração de Textos permite a transformação desse grande volume de dados textuais não estruturados em conhecimento útil, muitas vezes inovador para as organizações. Até pouco tempo esse fato não era visto como uma vantagem competitiva, ou como suporte à tomada de decisão, como indicativo de sucessos e fracassos. O seu uso permite extrair conhecimento a partir de dados

> textuais brutos (não estruturados), fornecendo elementos de suporte à gestão do conhecimento, que se refere ao modo de reorganizar como o conhecimento é criado, usado, compartilhado, armazenado e avaliado. Tecnologicamente, o apoio de Mineração de Textos à gestão do conhecimento se dá na transformação do conteúdo de repositórios de informação em conhecimento a ser analisado e compartilhado pela organização.

Segundo Scarinci (1997), um ponto muito importante que deve se frisar referente à extração de informação em textos é que a semântica de um arquivo textual não é representada por características superficiais como palavras individuais. Sendo assim, o desempenho e a eficácia de extração de informação é uma atividade que demanda uma enorme complexidade e que nem sempre traz resultados satisfatórios.

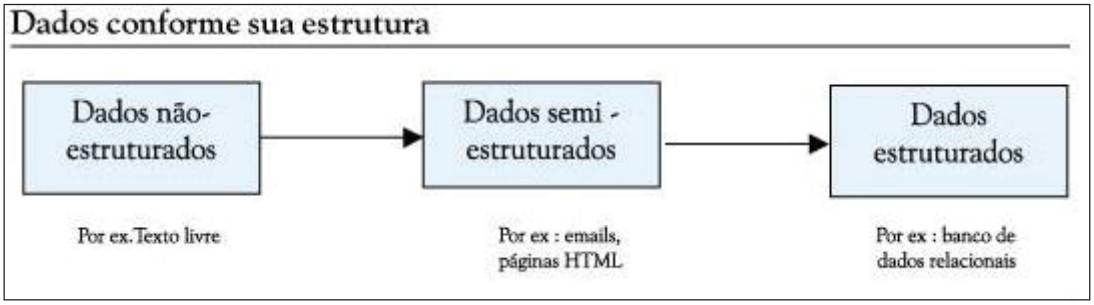
2.1.1 Tipos de dados

Para Miranda (1999, p.1), "dado é um conjunto de registros qualitativos ou quantitativos conhecidos que, quando organizado, agrupado, categorizado e padronizado adequadamente, transforma-se em informação". Pode-se dizer que são os registros de aspectos do fenômeno em estudo e que um pesquisador responsável pode-os capturar a partir de um determinado critério. Os dados em si são anotações diretas de resultados extraídos sobre observações, com pouca elaboração ou tratamento. Uma vez coletados, são compreendidos como um reflexo dos acontecimentos concretos de razoável confiabilidade.

Miranda (1999) afirma que a informação é o resultado da estruturação, transformação ou analise de dados, de forma lógica e concisa em seus resultados. É a tradução do que o conjunto de dados parece indicar.

Almeida (2002) facilita a representação dos tipos de dados, através da figura 1:

Figura 1 – Representação dos tipos de dados



Fonte: Almeida (2002; p.8).

Almeida (2002) informa que existem diferentes formas de distinguir dados em bases eletrônicas em que é possível ser interpretado.

> As páginas da Internet são consideradas dados semiestruturados, de caráter intermediário, ou seja, que apresentam alguma estrutura. Têm-se ainda os dados estruturados, como, por exemplo, aqueles presentes nos bancos de dados relacionais, além dos não estruturados, como, por exemplo, o texto livre (ALMEIDA, 2002, p.8).

A seguir, é descrito um pouco mais sobre cada tipo de dado e suas definições conforme citado anteriormente.

2.1.1.1 Estruturado

Para que um texto seja considerado estruturado, é necessário que o mesmo apresente uma regularidade no formato de apresentação das informações. Isso faz com que facilite as capturas realizadas por sistemas de informação, permitindo assim que os elementos sejam identificados com base em regras uniformes, podendo ser, por exemplo, marcadores textuais como os delimitadores de texto, ou a ordem de apresentação dos elementos. (ÁLVAREZ, 2007).

Almeida (2002) afirma que os dados estruturados são caracterizados por organizar suas instancias em regras bem definidas. Isso proporciona, através da aplicação de filtros e consultas, agrupar as informações e extrair dados relevantes das mesmas.

2.1.1.2 Semiestruturado

Almeida (2002) afirma que os dados semiestruturados representam uma grande parcela das informações disponíveis na internet, entretanto, esses dados nem sempre podem coincidir com uma base de conhecimento de determinada organização, já que cada um possui um padrão de estrutura diferente, que nem sempre é possível de manipular.

Os dados semiestruturados são denominados desta maneira, pois não estão em forma de texto bruto, que exige um grande processamento e possuem alguma estrutura. Entretanto, esse tipo de dado também não está totalmente estruturado, como em um banco de dados relacional, por exemplo. (ALMEIDA, 2002).

Os dados semiestruturados têm como característica a habilidade de variação na sua formulação e estrutura, os quais podem se adequar a diversas situações a padrões.

2.1.1.3 Não estruturado

Os dados não estruturados ou livres são conhecidos por não possuir nenhuma regularidade ou padrão na sua apresentação. Dessa forma, torna-se uma tarefa muito mais complexa, a extração de informações contidas no texto, a não ser que tenha um conhecimento linguístico sobre o dado que está sendo apresentado. (ÁLVAREZ, 2007).

Wives (2002) afirma que esse tipo de informação textual não é tratado pelas ferramentas tradicionais de descoberta de conhecimento, pois possuem características que tornam a análise e o tratamento desses dados muito mais complexos e trabalhosos.

Wives (2002) afirma que, para aplicar as etapas do processo de extração corretamente, são necessárias técnicas e ferramentas computacionais desenvolvidas especificamente para tratar esse tipo de informação não estruturada.

Essas técnicas de recuperação de informação de forma não estruturada, são conhecidas como descoberta de conhecimento em texto (*Knowledge Discovery from Text – KDT*).

A seguir, são apresentadas, algumas dessas técnicas de extração de informação.

2.1.2 Técnicas para extração de informação

As técnicas de extração de informação têm como principal objetivo minimizar o processamento de dados e dar mais agilidade à busca pela informação procurada.

Suponhamos que se deseja localizar o nome de uma pessoa em um banco de dados textual, de forma bruta, seria necessário realizar uma análise caractere por caractere do texto, comparando com o nome da pessoa, até encontrar a sequência correta, que se encaixe perfeitamente com o que está sendo procurado. Wives (1997) afirma que esse tipo de busca não é conveniente, pois exige muito processamento e acaba sendo uma tarefa muito demorada. Para isso, é preciso ter alguma maneira que torne esse tipo de consulta mais rápida e eficiente, desta forma, existem técnicas de processamento de linguagem natural, que facilitam esse processo. (WIVES, 1997).

Nesse mesmo contexto, Alvarez (2007) informa que técnicas de processamento de linguagem natural têm sido amplamente utilizadas no processo de extração de informação de documentos semiestruturados e não estruturados. Tendo como objetivo do uso dessas técnicas no processo da extração de informação, tentar analisar e compreender textos em alguma linguagem natural, a fim de encontrar informações relevantes que possam ser utilizadas de alguma forma.

LOH (2001) apresenta algumas técnicas para extração de informação em linguagem natural, as quais serão apresentadas a seguir.

2.1.2.1 Extração

Lehnert (1994, apud LOH 2001) afirma que uma técnica clássica de processamento de dados é a de Extração de Informação, que possui como principal objetivo encontrar informações específicas dentro dos textos.

O objetivo da área de extração de informação não é o mesmo ao objetivo da área de processamento de linguagem natural, porque é mais focado e mais bem definido, visando a extrair tipos específicos de dados.

Loh (2001) afirma que a técnica de extração de informação procura converter dados não estruturados em informações que possam ser mais bem analisadas. Geralmente, os métodos utilizados para extração servem para atender aplicações específicas.

"A estratégia mais utilizada é analisar "tags" (marcas) nos textos que possam indicar a presença de um dado. Por exemplo, o termo "anos" pode indicar que o numeral que o precede é a idade de alguém". (LOH e outros, 2001, p.6).

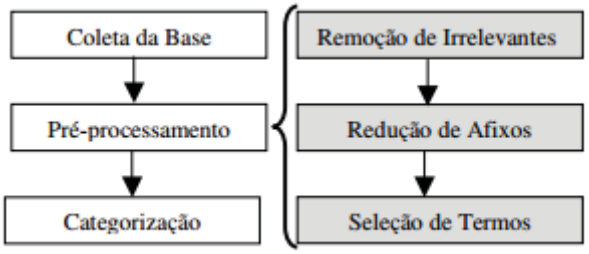
Através das "tags" é possível criar regras que possibilitam a extração de informações, mesmo que a fonte de dados pesquisada não possua estrutura alguma.

2.1.2.2 Categorização

Da Silva, Vieira e Osório (2005) definem a técnica de categorização como um processo de classificação de documentos em uma ou mais categorias predefinidas anteriormente. Esse processo de classificação pode ser utilizado em diversos contextos, desde a indexação automática de documentos, como em catálogos de recursos da Web ou em qualquer aplicação que exija organização de documentos ou seleção de mensagens.

A figura 2, a seguir, ilustra o processo realizado na categorização:

Figura 2 – Etapas do processo de categorização.



Fonte: da Silva, Vieira e Osório. (2005, p.1).

Da Silva, Vieira e Osório (2005) afirmam que o pré-processamento no processo de categorização de textos é considerado uma etapa essencial e muito custosa. Os textos originalmente não são estruturados e existe uma série de passos necessários para transformar os textos em um formato compatível para a extração de conhecimento.

Da Silva, Vieira e Osório (2005) informam que, na fase de categorização, os documentos são codificados e então apresentados a uma técnica de aprendizado de máquina.

Monard e Baranauskas (2003) definem o termo "aprendizado de máquina" como uma área da inteligência artificial, na qual o principal objetivo é o desenvolvimento de técnicas computacionais capazes de adquirir conhecimento de forma automática, ou seja, um sistema de computador que toma decisões baseado em experiências acumuladas através de problemas anteriores.

2.1.2.3 Análise de características ou descrição de conceitos (centroide)

A técnica de análise de características ou descrições de conceitos trata-se em apresentar uma lista com os conceitos principais de um único texto onde geralmente esses conceitos são termos ou expressões extraídos por análises estatísticas. É utilizada uma lista de termos próximos (antes e depois), os quais permitem a análise do conteúdo por quase frases (LOH, 2001).

Wives e Loh (2000) definem a técnica de centroide como uma técnica de extração de informação que procura valores de atributos dentro dos textos, tendo por objetivo extrair resumos de um texto ou de uma coleção, podendo ser uma visão geral ou as partes mais importantes ou mais interessantes. A técnica de listagem de conceitos chave, por sua vez, analisa uma coleção de textos em busca de características comuns (palavras, palavras-chave, temas), formando o que se convencionou chamar de centroide.

Wives e Loh (2000) informam que, assim como na técnica de análise de características ou descrição de conceitos apresentada na seção 2.1.2.3, a técnica da diferença também procura valores de atributos dentro de textos, porém sendo feito da forma inversa, isto é, descobrindo diferenças, comparando textos ou coleções.

2.1.2.4 Análise linguística

Segundo Loh (2001) define a abordagem por análise linguística como:

> A abordagem por análise linguística procura descobrir informações e regras analisando sentenças da linguagem a nível léxico, morfológico, sintático e semântico. Analisando padrões sintáticos (tags), as técnicas permitem descobrir generalizações escondidas, inferências de relações de coerência em textos (por exemplo, causa e efeito), relações de tempo e relações conceituais (definições, exemplos, partições e composição) através de *tags* no texto. (LOH, 2001, p.6).

Para Teline (2004), a extração por análise linguística de textos é realizada através de procedimentos computacionais, sendo realizadas em duas fases:

• pré-processamento linguístico e anotação automática de texto de dicionários online.

• consulta e a extração de informações relevantes para a execução de uma tarefa específica.

Dessa forma, tais consultas podem apresentar um melhor rendimento, e desempenho das informações identificadas extraídas.

Teline (2004) afirma que as etapas de análise linguística e anotações semânticas geralmente são constituídas pelas seguintes fases:

- tokenização: identificação de palavras e limites das sentenças;
- análise morfossintática: identificação de categorias gramaticais, características morfossintáticas e distribucionais;
- etiquetação part-of-speech: eliminação da ambiguidade de hipóteses da análise morfossintática e anotação das hipóteses mais prováveis no texto;
- lematização: identificação de candidatos a lema com base nos resultados da análise morfossintática e da etiquetação part-of-speech.

A utilização das técnicas de análise linguística, se bem formulada e elaborada, tem a capacidade de extrair a informação desejada, mesmo que o conteúdo pesquisado esteja em forma de texto bruto. As fases para formulação da análise semântica ajudam a identificar o conteúdo desejado, e descartar tudo aquilo que não se encaixa na regra elaborada.

2.1.2.5 Resumos ou sumarização

Loh (2001) afirma que a abordagem de descoberta por sumarização ou resumos utiliza as técnicas dos já apresentados anteriormente, entretanto, é mais voltada à produção do resumo ou sumário. Segundo SparckJones e Willet (1997, apud LOH 2001), sumarização é a abstração das partes mais importantes do conteúdo do texto.

Rino e Seno (2006) definem resumo ou sumarização automática de textos como sendo uma aplicação derivada do processamento de linguagem natural, esta técnica se restringe a duas metodologias sendo elas:

- **Baseada em informações linguísticas:** também conhecida como abordagem profunda e rica em conhecimento, nesta metodologia adotam-se modelos linguísticos ou discursivos que remetem à área fundamental de estudo das línguas naturais.

- **Baseada em informações estatísticas ou empíricas:** também conhecida como abordagem superficial ou pobre em conhecimento, nesta metodologia adotam-se modelos empíricos, matemáticos ou estatísticos, que são usados visando, sobretudo, ganhos computacionais, embora visem implicitamente a modelagem linguística.

Ser rica ou pobre em conhecimento, nesse contexto, significa diferenciar o tipo de conhecimento incorporado ao sistema, que lhe servirá de base para decisões de processamento. (RINO; SENO, 2006).

2.1.2.6 Associação entre textos

Loh (2001) informa que a descoberta por associação entre textos procura relacionar descobertas presentes em vários textos diferentes.

A técnica de associação (ou correlação) descobre relações de dependência entre textos ou características dos textos. (LOH; WIVES; OLIVEIRA, 2000).

Foram feitas descobertas na área médica, relacionando textos que não se referenciam e que aparentemente não continham assuntos comuns. Dessa forma é apresentada uma ferramenta que analisa diversos artigos sobre um mesmo evento e cria um resumo único em linguagem natural. São extraídas informações de partes dos textos e analisadas para encontrar similaridades e diferenças de informações. (LOH, 2001).

O objetivo dessa técnica é encontrar dependências entre atributos ou valores através da análise de probabilidades condicionais. Em geral, os resultados são apresentados na forma de regras X → Y, que significa que "se X está presente, então Y tem chances de estar presente também". O primeiro elemento X pode ser uma combinação de atributos ou valores, formando assim regras mais complexas. (LOH; GARIN; 2001).

Os autores apresentam duas regras internas para a aplicação da associação entre textos, apresentadas a seguir:

- **Análise de séries temporais:** Esta técnica procura encontrar padrões na repetição seguida de valores. Por exemplo, analisando-se as ações de uma empresa na bolsa de valores, pode-se notar que depois de tantos meses subindo, o valor das ações diminui em tantos pontos percentuais.

- **Evolução ou sequência de tempo:** As técnicas de evolução ou sequência de tempo buscam descobrir regras de associação ou correlação entre eventos ocorridos em momentos diferentes. Por exemplo, uma loja pode identificar que clientes que compram um sapato voltam depois de um mês para comprar uma camisa.

Furtado (2004) informa que a técnica de associação entre texto realiza uma análise de diversos documentos sobre um mesmo evento e extrai informações de partes dos textos,

por técnicas tradicionais de extração de informação, as quais são estruturadas em *slots* (pares atributo-valor, representando internamente conceitos), que são analisados para encontrar similaridades e diferenças de informações.

Existem poucas ferramentas automáticas e concretas para esse tipo de abordagem. O que geralmente acontece é haver técnicas sistemáticas, empregadas por pessoas, mas que exigem ainda muita interpretação humana.

2.1.2.7 Clustering

Loh, em 2001, afirma que a descoberta por agrupamento, podendo ser chamada também de *clustering,* tem como objetivo separar automaticamente elementos em grupos por alguma afinidade ou similaridade. Essa técnica de agrupamento é diferente da classificação, pois a primeira objetiva criar as classes através da organização dos elementos, enquanto que a segunda procura alocar elementos em classes já pré-definidas.

Essa técnica de agrupamento auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) nas classes. (LOH, 2001).

2.2 EXPRESSÕES REGULARES

Segundo Goyvaerts e Levithan (2011), uma expressão regular é um tipo específico de texto-padrão que pode ser utilizado em muitos aplicativos modernos e em linguagens de programação. As expressões regulares possibilitam verificar se a entrada de dados se encaixa no padrão de texto, para encontrar um texto que corresponda a um padrão dentro de um conjunto maior de textos, para substituir o texto padrão por outro ou reorganizar bits de texto correspondentes e para dividir um bloco de texto em uma lista de subtextos.

> Uma expressão regular é uma composição de símbolos, caracteres com funções especiais, que, agrupados entre si e com caracteres literais, formam uma sequência, uma expressão. Essa expressão é interpretada como uma regra, que indicará sucesso se uma entrada de dados qualquer casar com essa regra, ou seja, obedecer exatamente a todas as suas condições, resumindo é um método formal de se especificar um padrão de texto. (JARGAS, 2009, p.13).

Para Jargas (2009), as expressões regulares são uma grande aliada no mundo tecnológico informatizado. Quando mais refinada for a sua formulação, mais preciso e rápido é o resultado.

Apesar de a maioria das linguagens de programação, programas e editores de texto mais utilizado possuir esse recurso, são poucos os que o dominam, principalmente pelo fato de a documentação sobre o assunto, quando existente, ser pouco didático. (JARGAS, 2009).

Ferreira (2005) afirma que uma expressão regular é uma notação algébrica que caracteriza um conjunto de caracteres e esse conjunto pode ser usado para procurar e identificar elementos num texto, assim como definir uma linguagem de modo formal.

No Quadro 1 são apresentados exemplos de uso de expressões regulares, aplicadas em uma frase:

Quadro 1 – Exemplos de expressões regulares.

Padrão a identificar	**Aplicação prática**	**Comentários**
/Helder/	Bom dia Helder!	encontrou 1 ocorrência
/e+/	Bom dia Helder!	encontrou 1 ocorrência
	Preencheu tudo?	encontrou 1 ocorrência
/E/	Bom dia Helder!	não existe E maiúsculo

Fonte: FERREIRA (2005; p 103).

Ferreira (2005) afirma que o uso de expressões regulares requer a existência de um padrão que queremos identificar e um texto onde procurar.

Dessa forma, são aplicadas regras para busca dessas expressões dentro de arquivos textuais. A seguir é apresentada uma figura demonstrando o uso de regras para definição da busca aplicadas em um contexto, utilizando metacaracteres, que servem como filtro para aplicação das expressões regulares, por exemplo, o uso do caractere '^', significa que a expressão pesquisada deve estar no início da linha, conforme apresentado no Quadro 2:

Quadro 2 – Uso de operadores em expressões regulares.

Meta-caracter	Descrição	Exemplo	Aplicação prática
^	início de linha	/^Bom/	Bom dia Helder!
$	fim de linha	/Bom$/	Hoje é um dia bom
|	ou	/a|b/	Hoje é um dia bom.
			Hoje é um bom dia.
.	qualquer caracter	/Bo./	Bom dia Helder!
+	um ou mais	/ra.+/	Coisa rara não é?
*	zero ou mais	/carro*/	Bom dia Helder!
		/carro*/	Belo carro!
?	um opcional	/B(o)?m/	Bom carro.
		/B(o)?m/	É um Bmw?
(...)	captura padrão	/B(o|a)m?/	Bom dia. Captura: "o".
		/B(o|a)m?/	Belo barco. Captura: "a".

Fonte: FERREIRA (2005; p 104).

O Quadro 3 apresenta o exemplo de como negar os conteúdos de uma classe de caracteres utilizado em uma expressão regular com recurso ao metacaractere "^". (FERREIRA, 2005).

Quadro 3 – Exemplo negação em expressões regulares.

Padrão a identificar	Aplicação prática	Comentários
/[^e]+/	Boneca!	ignora o "e" .

Fonte: FERREIRA (2005; p 104).

As expressões regulares são consideradas essenciais para que os algoritmos de processamento de texto sejam potentes, eficientes e flexíveis. A aplicação de expressões regulares permite a descrição e

> análise de texto, assim como, remoção, adição e manipulação de blocos de frases. (FERREIRA, 2005, p 104).

O uso de expressões regulares facilita a procura por elementos que se encontram em arquivos textuais, possibilitando sua extração de forma rápida e limpa. Para isso, basta que as expressões regulares estejam bem formuladas.

2.3 FORMATOS DE REFERÊNCIAS BIBLIOGRÁFICAS

Na elaboração de trabalhos acadêmicos, é indispensável a consulta de diversas fontes de informação que devem ser devidamente identificadas de forma uniformizada e de acordo com a aplicação de normas apropriadas. As citações apoiam uma hipótese, sustentam uma ideia ou ilustram um raciocínio, oferecendo ao leitor o respaldo necessário para que ele possa comprovar a veracidade das informações fornecidas e possibilitar o seu aprofundamento. (INSTITUTO DE RELAÇÕES INTERNACIONAIS – USP, 2012).

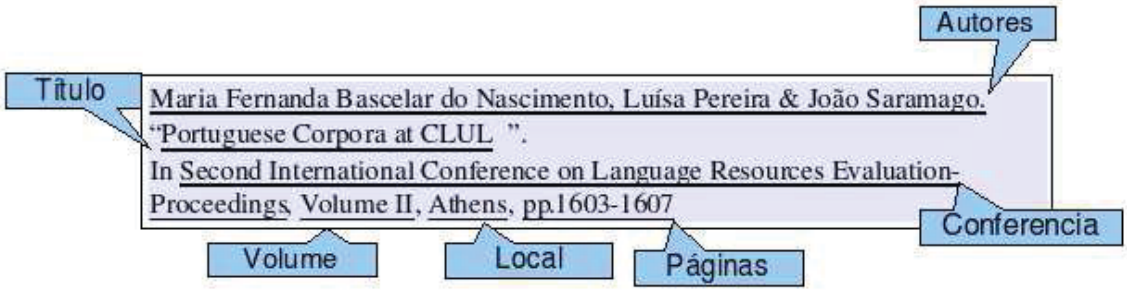
Segundo o Instituto de Relações Internacionais da USP (2012), o formato da referência solicitado pela instituição ou editora responsável por uma publicação é especificado nas instruções para autores. Em geral, para o campo de Relações Internacionais, o formato ABNT é indicado para as publicações nacionais e o formato Chicago para as publicações internacionais.

Existem diversas outras normas nacionais e internacionais para a elaboração de referências bibliográficas, dentre elas, podemos citar: Harvard, APA, NP405.

Segundo a Associação Brasileira de Normas Técnicas (ABNT), referência bibliográfica é o conjunto de elementos que permite a identificação, no todo ou em parte, de documentos impressos ou registros em diversos tipos de materiais. Deve obedecer a NBR 6023, Informação e documentação referências – Elaboração (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2002, p. 2).

Partes da referência, como autor, título, ano, nome da conferência ou nome da revista são exemplos de elementos bibliográficos. Na figura 3, é possível ver os elementos bibliográficos destacados numa referência bibliográfica.

Figura 3 - Exemplo de uma referência bibliográfica com os elementos bibliográficos.



Fonte: CABRAL (2005; p 27).

> Existe uma diversidade de fontes de informação: documentos impressos, manuscritos, registros audiovisuais, sonoros, magnéticos e eletrônicos e outros que podem ser apresentados on-line ou em diferentes suportes como: CD-ROM, disquetes, fitas magnéticas. A referência de documentos não convencionais deve ser acrescida de dados específicos que possibilitem sua localização e recuperação. (CABRAL, 2005, p.27).

Cabral (2005) indica dois tipos de formatos referências que são apresentados a seguir:

2.3.1 Texto simples

Cabral afirma que, em texto simples, as informações são apresentadas sem qualquer separador específico. Uma referência bibliográfica pode ser representada em vários estilos, alterando a disposição e apresentação dos elementos bibliográficos no texto. Diferentes formas de representação gráfica constituem diferentes estilos bibliográficos. (CABRAL, 2005).

2.3.2 Estruturado

Cabral (2005) afirma que outro modo de representar referências é num formato estruturado onde cada elemento bibliográfico está devidamente identificado e delimitado. Esta forma de representação será designada de formato bibliográfico. Existem vários formatos bibliográficos, mas são distintos, facilmente reconhecíveis, e o seu objetivo é poderem ser processados por programas com uma certa facilidade.

Cabral (2005) afirma que são precisamente as diferenças entre cada uma dessas representações que justificam o seu uso. O estilo bibliográfico tem como finalidade ser lido por seres humanos, necessita ser "legível", ajustando-se às necessidades da publicação que representam ou do domínio a que pertencem, exibindo ou ocultando diferentes elementos bibliográficos. Os formatos bibliográficos, por outro lado, foram desenhados para serem legíveis por programas, de forma a serem arquivado ou para produzir representações num determinado estilo bibliográfico vital que se possa distinguir sem ambiguidade todas as partes da referência. É possível fazer a transformação de qualquer formato para qualquer estilo bibliográfico. No entanto, o processo inverso não tem necessariamente de ocorrer. (CABRAL, 2005).

2.4 INDICADORES DE PUBLICAÇÕES ACADÊMICAS

Silva e outros (2009) definem a geração de indicadores de publicações acadêmicas, também é conhecido como análise bibliométrica, como de uma ferramenta na qual o principal objetivo é medir a produção científica, ou seja, medir o fator de impacto do documento, atualidade das referências, importância dos autores utilizados, etc.

> Fazer o levantamento do inventário das atividades científicas, nos mais diversos campos do conhecimento, implica em uma busca criteriosa nas publicações, pois o homem busca apresentar constantemente novos conhecimentos, fazendo com que as informações circulem e se disseminem por todas as partes do mundo. (SILVA; FILHO E PINTO, 2009, p.3).

Mugnaini e outros (2004) afirmam que bibliometria, como área de estudo da ciência da informação, tem um papel relevante na análise da produção científica, uma vez que seus indicadores retratam o grau de desenvolvimento de uma área do conhecimento de um campo científico ou de saber.

A análise dos indicadores bibliométricos permite, por meio de análise estatística, quantificar a produção científica e técnica.

Há, por parte de autores, a ideia de que a avaliação da produtividade científica, por exemplo, deve ser um dos elementos principais para o estabelecimento e acompanhamento de uma política nacional de ensino e pesquisa, uma vez que permite um diagnóstico das reais potencialidades de determinados grupos e/ou instituições (VANTI, 2002).

Como informa Alvarenga (1998), elementos textuais referentes a documentos científicos, como monografias e artigos de periódicos, compõem-se por variáveis abordadas nos estudos da bibliometria. Desta forma, é possível alcançar resultados que refletem quantitativamente o conhecimento abordado nos textos, apontando campos como produtividade de autores, os autores que constituem as frentes da pesquisa. Nesse sentido, o potencial de dados gerados pela bibliometria se apresenta como insumos valiosos para o desenvolvimento de estudos arqueológicos e epistemológicos regionais, ou seja, dos campos específicos do saber.

Vanti (2002), entretanto, questiona de que maneira é possível fazer este diagnóstico. Uma das possibilidades consiste na utilização de métodos que permitam medir a produtividade dos pesquisadores, grupos ou instituições de pesquisa. Para tanto, torna-se fundamental o uso de técnicas específicas de avaliação que podem ser quantitativas ou qualitativas, ou mesmo uma combinação entre ambas.

Segundo Vanti (2002), as técnicas quantitativas de avaliação podem ser subdivididas em bibliometria, cienciometria, informetria e, mais recentemente, webometria. Estas técnicas têm funções semelhantes, entretanto cada uma propõe medir o conhecimento cientifico e o fluxo de informação em enfoques diferentes.

De acordo com as palavras de Tague-Sutckiffe (apud, VANTI, 2002), pode-se definir a bibliometria como: “[...] o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. A bibliometria desenvolve padrões e modelos matemáticos para medir esses processos, usando seus resultados para elaborar previsões e apoiar tomadas de decisões”.

O mesmo autor define cienciometria, outro segmento área da bibliometria, como sendo o estudo dos aspectos quantitativos da ciência, enquanto uma disciplina ou atividade econômica, ou seja, um segmento da sociologia da ciência, sendo aplicada no desenvolvimento de políticas científicas que envolvem estudos quantitativos das atividades científicas, incluindo a publicação.

> A cienciometria estuda, por meio de indicadores quantitativos, uma determinada disciplina da ciência. Estes indicadores quantitativos são utilizados dentro de uma área do conhecimento, por exemplo, mediante a análise de publicações, com aplicação no desenvolvimento de políticas científicas. Tenta medir os incrementos de produção e produtividade de uma disciplina, de um grupo de pesquisadores de uma área, a fim de delinear o crescimento de determinado ramo do conhecimento. (VANTI, 2002, p. 154).

Vanti (2002) diz que a informetria, diferentemente da cienciometria e da bibliometria, não se limita apenas à informação registrada. Neste tipo de técnica, é possível analisar os processos de comunicação informal, como a falada, e se dedicar a pesquisar o uso e a necessidade de informações dos grupos sociais desfavorecidos e não os intelectuais.

> Informetria é o estudo dos aspectos quantitativos da informação em qualquer formato, e não apenas registros catalográficos ou bibliografias, referente a qualquer grupo social, e não apenas aos cientistas. A informetria pode incorporar, utilizar e ampliar os muitos estudos de avaliação da informação que estão fora dos limites da bibliometria e cienciometria. (TAGUE-SUTCKIFFE, apud VANTI; 2002, p. 155).

Vanti (2002) afirma que as definições teóricas podem ajudar na compreensão do que pode ser cada um dos métodos mencionados anteriormente, entretanto, para facilitar o entendimento, é possivel tentar associar esses métodos à utilização de aplicações concretas. Em termos genéricos, estas são algumas possibilidades de aplicação das técnicas bibliométricas, cienciométricas e informétricas:

- identificar as tendências e o crescimento do conhecimento em uma área;
- identificar as revistas do núcleo de uma disciplina;
- mensurar a cobertura das revistas secundárias;
- identificar os usuários de uma disciplina;
- prever as tendências de publicação;
- estudar a dispersão e a obsolescência da literatura científica;
- prever a produtividade de autores individuais, organizações e países;

- medir o grau e padrões de colaboração entre autores;
- analisar os processos de citação e co-citação;
- determinar o desempenho dos sistemas de recuperação da informação;
- avaliar os aspectos estatísticos da linguagem, das palavras e das frases;
- avaliar a circulação e uso de documentos em um centro de documentação;
- medir o crescimento de determinadas áreas e o surgimento de novos temas.

Vanti (2002) afirma que os índices bibliométricos são utilizados, também, para avaliar a produtividade e a qualidade da pesquisa de cientistas, isso se da por meio da medição com base nos números de publicações e citações de diversos pesquisadores.

Como se pode analisar, os indicadores de publicações acadêmicas são utilizados em diversas maneiras e para diversos objetivos, tendo grande utilidade na área da bibliometria.

Nesta sessão se pôde entender um pouco sobre a geração de indicadores, e sua utilidade no ambiente acadêmico.

No próximo capitulo é apresentado os métodos utilizados para a pesquisa, voltado ao âmbito do desenvolvimento da proposta de solução.

3 MÉTODO

Neste capitulo, é apresentada a caracterização do tipo de pesquisa na qual o trabalho se enquadra, apresentado detalhadamente cada um dos tipos. Também são apresentadas as etapas metodológicas para o desenvolvimento de todo o trabalho, além de um modelo proposto para a solução do problema.

3.1 CARATERIZAÇÃO DO TIPO DE PESQUISA

Toda pesquisa científica necessita definir seu objeto de estudo e, a partir daí, construir um processo de investigação, delimitando o universo que será estudado (KAUARK; MANHÃES; MEDEIROS; 2010).

Segundo Kauark e outros (2010), a importância de conhecer os tipos de pesquisas existentes ajuda na definição dos instrumentos e procedimentos necessários que o pesquisador precisará utilizar para o planejamento da sua pesquisa. Dessa forma, é necessário que o pesquisador saiba usar os instrumentos adequados para que possa achar a solução do problema levantado pela sua pesquisa.

As seções seguintes apresentam as características do tipo de pesquisa utilizadas para o desenvolvimento deste trabalho. É descrita a pesquisa bibliográfica, exploratória e qualitativa.

3.1.1 Bibliográfica

Kauark e outros (2010) definem a pesquisa bibliográfica como sendo um método de pesquisa utilizado para desenvolvimento a partir de material já publicado por autores da área, constituído principalmente de livros, artigos de periódicos e material disponibilizado na Internet.

Leonel e Motta (2007) definem esse tipo de pesquisa como sendo aquela que se desenvolve tentando explicar um problema a partir das teorias publicadas em diversos tipos de fontes: livros, artigos, manuais, enciclopédias, anais, meios eletrônicos, etc, tendo como objetivo conhecer e analisar as principais contribuições teóricas sobre um determinado tema ou assunto.

Os autores apontam que, para a escolha do tema para realização de uma pesquisa bibliográfica, deve-se levar em consideração o interesse pelo assunto, a disponibilidade bibliográfica especializada já realizada por autores da área, além da familiaridade com o assunto.

3.1.2 Exploratória

Vieira (2002) aponta que a pesquisa exploratória tem o objetivo de deixar o pesquisador familiarizado com o assunto. Este esforço tem como meta deixar o problema que antes parecia complexo, muito mais explicito, facilitando a pesquisa sobre o conteúdo abordado e conhecendo bastante sobre o assunto.

> A pesquisa exploratória é usada em casos nos quais é necessário definir o problema com maior precisão e identificar cursos relevantes de ação ou obter dados adicionais antes que se possa desenvolver uma abordagem. Como o nome sugere, a pesquisa exploratória procura explorar um problema ou uma situação para prover critérios e compreensão. (VIEIRA, 2002, p.7).

A caracterização do estudo como pesquisa exploratória normalmente ocorre quando há pouco conhecimento sobre a temática a ser abordada. Por meio do estudo exploratório, busca-se conhecer com maior profundidade o assunto de modo a torná-lo mais claro ou construir questões importantes para a condução da pesquisa. (RAUPP; BEUREN, 2003).

3.1.3 Quantitativa

Segundo Kauark e outros (2010) explicam que, a pesquisa quantitativa é considerada como aquela que pode ser quantificável, ou seja, possibilita traduzir em números, opiniões e informações para classifica-las e analisá-las.

"A pesquisa quantitativa requer o uso de recursos e de técnicas estatísticas (percentagem, média, moda, mediana, desvio-padrão, coeficiente de correlação, análise de regressão)". (KAUARK e outros, 2010, p.27).

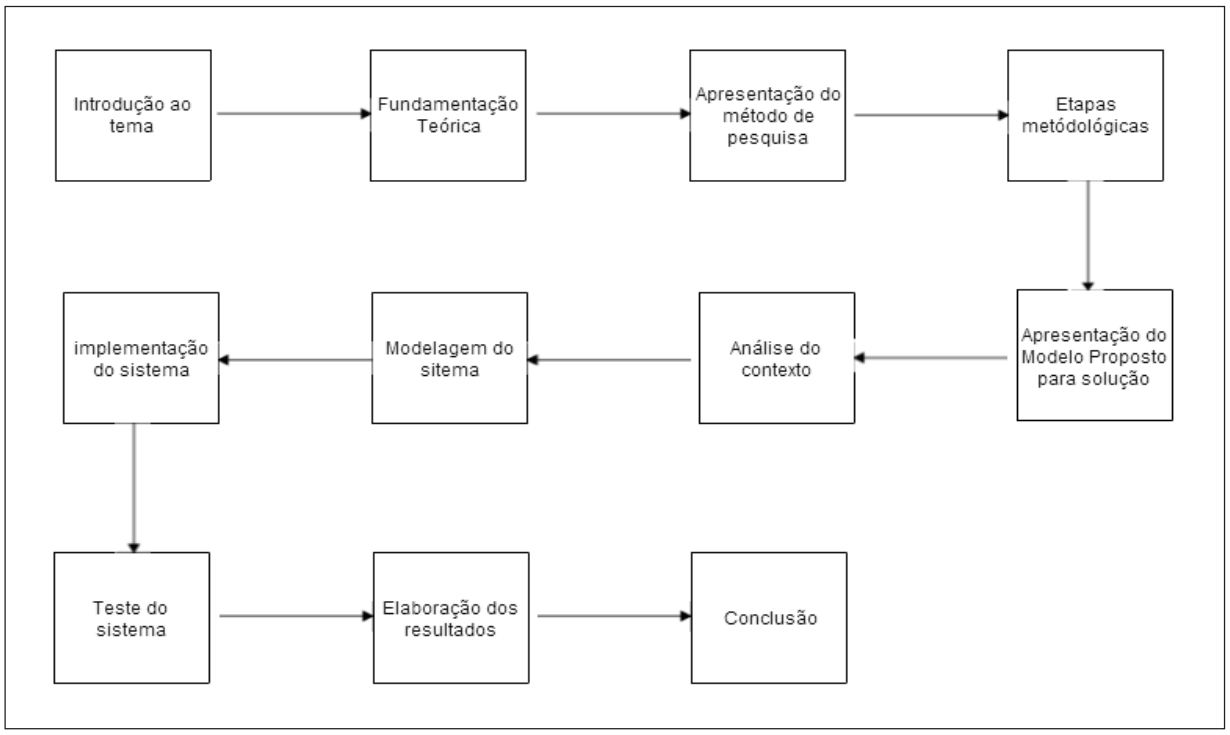
Ao utilizar uma abordagem de pesquisa quantitativa, é necessário utilizar sempre o recurso das representações gráficas, que geralmente são representadas por tabelas, quadros e gráficos com o objetivo de facilitar a análise e a interpretação dos dados (LEONEL; MOTTA, 2007).

Neste trabalho em questão, a pesquisa quantitativa se dará pelo nível de acerto e integridade das informações extraídas de documentos textuais, comparando o resultado apresentado pelo sistema desenvolvido, com a real informação que é adicionada ao mesmo.

3.2 ETAPAS METODOLÓGICAS

Neste tópico são apresentadas todas as etapas necessárias para desenvolvimento do presente trabalho, contemplando a implementação da solução, a Figura 4 demonstra as etapas em uma forma de diagrama para melhor entendimento, a seguir, é detalhada cada uma dessas etapas.

Figura 4 – Etapas Metodológicas para desenvolvimento do trabalho.



Fonte: Autor, 2015.

Conforme apresentado no diagrama, ao longo deste trabalho, serão apresentadas 11 etapas, sendo elas:

- introdução ao tema: Nesta etapa, é apresentada uma introdução sobre o contexto do trabalho, o problema da pesquisa e a justificativa para solução;
- fundamentação teórica: Etapa em que é apresentada toda a fundamentação teórica de diversos autores da área, para apresentar o conteúdo da pesquisa;
- apresentação do método de pesquisa: Nesta etapa, são apresentados os tipos de pesquisa em que o presente trabalho se enquadra;
- etapas metodológicas: Esta trata do presente tópico, em que são apresentadas as etapas para elaboração do trabalho;
- apresentação do modelo proposto para solução: Nesta etapa, é apresentado um macromodelo, representando as principais atividades realizadas pelo sistema, para conseguir chegar ao objetivo do trabalho;
- analise do contexto: Etapa em que é realizada a análise de requisitos para realização do sistema;

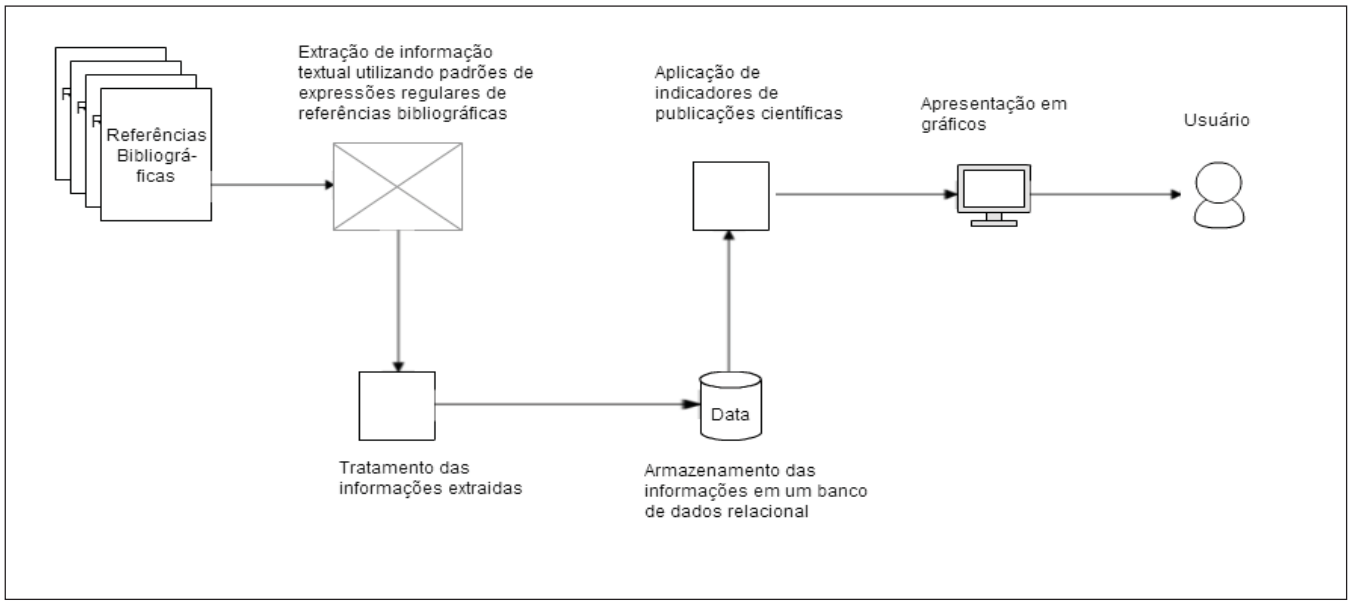
- modelagem do sistema: Nesta etapa, é apresentado o modelo Iconix, para desenvolvimento do sistema, contendo diagramas de classe, caso de uso, sequência e robustez;
- implementação do sistema: Etapa em que é desenvolvido o código fonte de extração de informações de texto, fazendo o uso de expressões regulares aplicadas sobre referências bibliográficas nas normas da ABNT de elaboração de referências, armazenamento estas informações em banco de dados, e gerando indicadores para apresentar resultados de buscas nos dados extraídos;
- teste do sistema: Nesta etapa, são realizados testes para disponibilizar a execução do sistema sem problemas;
- elaboração dos resultados: Etapa em que são apresentados os resultados obtidos com a elaboração do sistema, e resultados obtidos;
- conclusão: Nesta etapa, é apresentada a conclusão do trabalho e a realização de trabalhos futuros.

Na próxima seção, é apresentado um modelo proposto para a solução do trabalho.

3.3 PROPOSTA DE SOLUÇÃO

Nesta seção, é apresentado o detalhamento do modelo elaborado para proposta de solução, a Figura 5 demonstra o modelo proposto em forma de diagrama.

Figura 5- Demonstração do modelo proposto para a solução.



Fonte: Autor, 2015.

O modelo proposto descreve-se da seguinte forma:

O sistema realiza a extração de dados em texto sobre referências bibliográficas, estando estas em formato descrito pelas normas da ABNT de elaboração de referências, utilizando uma expressão regular na qual as referências estarão compostas. Após realizar a extração, o sistema irá tratar as informações extraídas, separando cada informação contida (ano de publicação, autor, local, titulo, etc.) e as armazenará em um banco de dados relacional.

Após toda a informação recuperada estiver no banco de dados, o sistema irá realizar a aplicação de indicadores de publicações acadêmicas.

Esses indicadores serão apresentados de forma gráfica, para melhor entendimento, e captura da informação pelo usuário final.

Para a avaliação do modelo proposto, será apresentado o funcionamento de um protótipo que irá realizar a leitura das referencias bibliográficas de um documento em texto, e apresentar as informações recuperadas sobre este mesmo documento, além de disponibilizar o protótipo para avaliação de profissionais da área.

3.4 DELIMITAÇÃO DA PESQUISA

A extração de informação do modelo proposto utiliza padrões de referências bibliográficas, para buscar as informações esperadas, nos locais esperados. Desta forma, o modelo não abrange todos os tipos de referências bibliográficas existentes, limitando-se apenas a as referências elaboradas no padrão da ABNT.

Como se sabe, referências bibliográficas são compostas por vários elementos (autores, editora, local, ano de publicação, etc.). Entretanto, o modelo proposto esta elaborado para trabalhar sob duas principais informações, sendo elas: autores e ano de publicação. Porém, o sistema está preparado para evoluir, e buscar, também, as demais informações contidas nas referências bibliográficas.

A geração de indicadores, geralmente, é realizada sobre uma grande massa de dados. Entretanto, o modelo não buscará um número excessivo de documentos, não possuindo, assim, um data warehouse, na geração de informações. E, sim, trabalhando apenas com as informações recuperadas das referências bibliográficas adicionadas ao sistema.

4 PROJETO DE SOLUÇÃO

Neste capitulo, é apresentado o projeto elaborado para solução do problema proposto. Primeiramente são demonstradas as técnicas utilizadas e, em seguida, o modelo piloto elaborado para desenvolvimento do projeto.

4.1 DEFINIÇÕES DE TÉCNICAS E METODOLOGIAS

Nesta seção, são apresentadas as técnicas e metodologias existentes que foram abordadas para aplicação da solução proposta e desenvolvimento do projeto.

4.1.1 Linguagem de Modelagem Unificada (UML)

BOOCH e outros (2006) afirmam que:

> A Linguagem de Modelagem Unificada (UML) é uma linguagem gráfica para visualização, especificação, construção e documentação de artefatos de sistemas complexos de software. A UML proporciona uma forma-padrão para a preparação de planos de arquitetura de projetos de sistemas, incluindo aspectos conceituais tais como processos de negócios e funções do sistema, alem de itens concretos como as classes escritas em determinada linguagem de programação, esquemas de bancos de dados e componentes de software reutilizáveis. (BOOCH; RUMBAUGH; JACOBSON, 2006, p.7).

Fowler (2005) afirma que a UML é um padrão relativamente aberto, controlado pela OMG (Object Management Group), um consórcio aberto de empresas. O OMG foi formado para estabelecer padrões que suportassem interoperabilidade, especificamente de sistemas orientados a objetos.

A UML é apenas uma linguagem e, portanto, é somente uma parte de um método para desenvolvimento de software. A UML é independente do processo, apesar de ser perfeitamente utilizada em processo orientado a casos de usos, centrado na arquitetura, iterativo e incremental (BOOCH; RUMBAUGH; JACOBSON, 2006).

BOOCH e outros (2006) apresentam a UML composta por diagramas, sendo esses, representações gráficas de um conjunto de elementos geralmente representados por itens (gráficos de vértices) e relacionamentos (arcos). Estes elementos são combinados para permitir a visualização de um sistema sob diferentes perspectivas. Em todos os sistemas, exceto os mais triviais, os diagramas representam uma visão parcial dos elementos que compõem o sistema. A seguir são apresentados os diagramas que compõem a linguagem de modelagem unificada (UML).

- Diagrama de classes
- Diagrama de objetos
- Diagrama de componentes
- Diagrama de estruturas compostas
- Diagrama de casos de uso
- Diagrama de sequencias
- Diagrama de comunicações
- Diagrama de gráficos de estados
- Diagrama de atividades
- Diagrama de implantação
- Diagrama de pacote
- Diagrama de temporização
- Diagrama de visão geral da interação

Nas próximas seções, são apresentadas as metodologias de desenvolvimento do sistema proposto e a tecnologia de orientação a objeto.

4.1.2 Orientação a Objetos (OO)

Kamienski (1996) afirma que é sempre muito custoso construir um sistema, geralmente, os prazos não são cumpridos e a manutenção é um grande problema para as empresas. Desta forma, o principal desavio para uma empresa ao desenvolver um software é construí-lo de forma rápida, barata e flexível.

"Os ambientes de desenvolvimento de software têm evoluído muito nos últimos anos para incorporar várias tecnologias num movimento sinergético em direção a obter melhores condições do desenvolvimento de sistemas." (KAMIENSKI, 1996, p.3).

Kamienski afirma que a orientação a objetos veio para facilitar o desenvolvimento de software, sendo esta uma metodologia de desenvolvimento que visa ao reaproveitamento de código e à modularidade dos sistemas.

4.1.3 ICONIX

Desenvolver um sistema de software, hoje em dia, é muito mais difícil do que se pode imaginar, os programas de computadores estão se tornando cada vez complexos e contendo cada vez mais informações para seus usuários. Por outro lado, os portais de ajudas à resolução de problemas também estão em forma crescente, entretanto, com os sistemas mais complexos, os problemas também são cada vez mais difíceis de ser resolvidos. Por este motivo, é muito importante que se tenha um planejamento e que se utilizem padrões existentes, para que facilite assim a manutenção nos programas de computadores. Isso pode ser realizado, utilizando um processo de desenvolvimento do software. (MAIA, 2005).

Desenvolvido pela ICONIX *Software Engineering*, o ICONIX é uma metodologia de desenvolvimento de software pura, prática e simples, mas também poderosa e com um componente de análise e representação dos problemas sólido e eficaz. (MAIA, 2005).

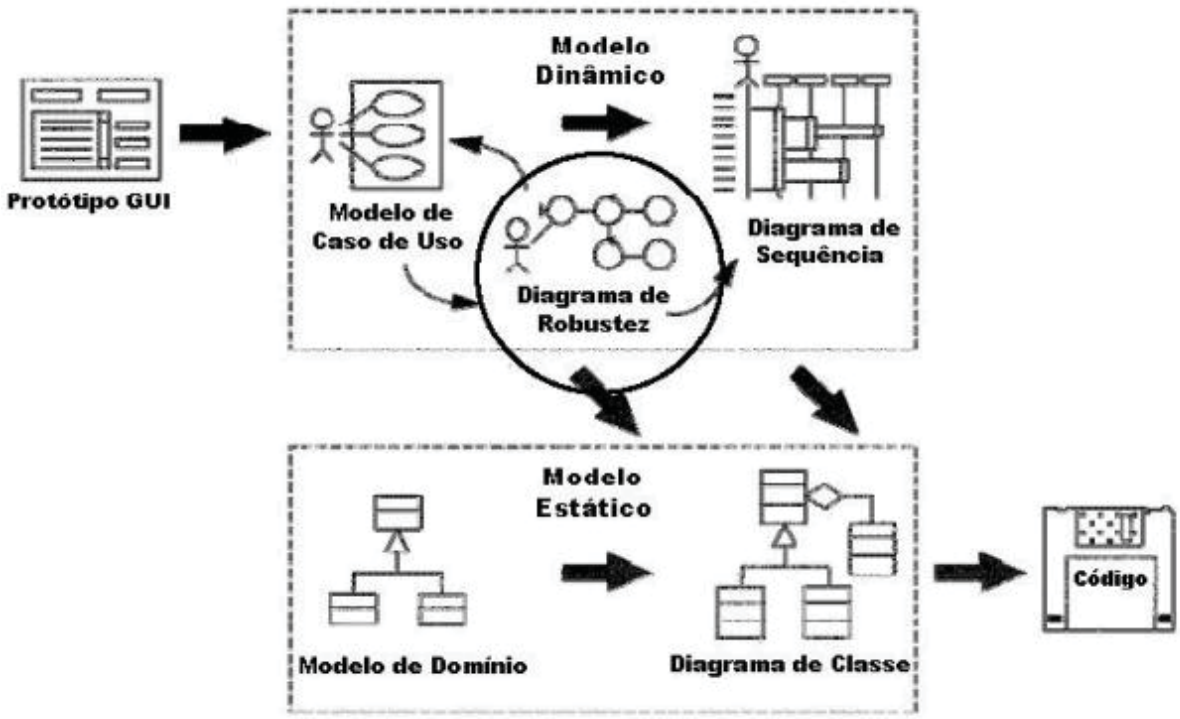
Maia (2005) afirma que essa metodologia também utiliza a linguagem de modelagem UML, entretanto, possui uma característica exclusiva, chamada de "Rastreabilidade dos Requisitos", que permite verificar em todas as fases se os requisitos do sistema estão sendo atendidos de forma completa.

O ICONIX é composto pelas seguintes principais fases que serão abordadas nas próximas sessões:

- Modelo de Domínio;
- Modelo de Caso de Uso;
- Análise Robusta;
- Diagrama de Sequência;
- Diagrama de Classe.

A Figura 6 apresenta a representação do modelo ICONIX, utilizando as fases citadas anteriormente.

Figura 6 – Representação do modelo ICONIX



Fonte: SILVA, 2007, p.4.

Maia (2005), em sua publicação sobre o processo de desenvolvimento ICONIX, afirma que o modelo é utilizado para desenvolver diagramas de caso de uso baseado em requisitos de usuários. Estes diagramas são a base para a realização de uma analise robusta para cada um dos casos de uso propostos e, com o resultado da analise robusta, é possível dar partida no diagrama que sequência, sendo que, após seu término, é possível modificar o modelo de domínio inserindo a ele os métodos e atributos descobertos no modelo de sequência.

Nesta seção, foi possível acompanhar o funcionamento do processo de desenvolvimento utilizando ICONIX. A próxima seção apresenta a parte prática, sendo esta a modelagem do sistema proposto que antecede seu desenvolvimento.

4.2 MODELAGEM DO SISTEMA PROPOSTO

Nesta seção do trabalho, é apresentada a modelagem do sistema proposto. Na primeira etapa, são apresentados os atores que irão compor o sistema. Em seguida, é apresentada a parte de requisitos, sendo composta por: requisitos funcionais, requisitos não funcionais e regras de negócio do sistema.

Seguindo a metodologia de desenvolvimento apresentada, são apresentados os modelos e diagramas que compõem a modelagem do sistema, sendo eles: diagrama de caso de uso, prototipação das telas do sistema, diagrama de sequência, diagrama de robustez e, por último, o diagrama de domínio que, com a evolução do processo, se tornará um diagrama de classes.

4.2.1 Atores

O sistema proposto contará com apenas um ator, chamado de Usuário, que pode fazer toda e qualquer operação no sistema, se registrado na plataforma.

4.2.2 Análise de Requisitos

Segundo Carvalho e outros (2001), a engenharia de requisitos é um dos pontos de maior interesse, sendo demonstrado pela indústria de software, trata-se de entender o que é desejado construir antes mesmo de começar a desenvolver.

O autor afirma que o tempo que é utilizado para o entendimento do problema é um excelente investimento, sendo que esse tempo certamente será recompensado depois, já que será desenvolvido com base no que já foi levantado, ao invés de iniciar o desenvolvimento sem um planejamento correto, fazendo com que necessite de tempo para rever os problemas e refazer partes do sistema desenvolvido.

> Os requisitos de software são a base a partir da qual a qualidade é medida. Desta forma, a falta de conformidade aos requisitos significa falta de qualidade. O modelo de avaliação de maturidade do processo de desenvolvimento CMM (Capability Maturity Model) considera o gerenciamento de requisitos como sendo uma das primeiras etapas para alcançar a maturidade organizacional, e para haver o gerenciamento é preciso que o processo de desenvolvimento de requisitos esteja implantado na empresa. (CARVALHO e outros, 2001, p.33).

Sendo assim, para que seja possível ter um bom gerenciamento do projeto, é de extrema importância que os requisitos tenham sido definidos de forma correta, sendo que é importante que seja possível o desenvolvimento dos requisitos propostos. (CARVALHO e outros, 2001).

Existem algumas classificações para os requisitos seguindo alguma forma de categorização que podem ser definidas conforme as práticas de cada organização de software. Na maioria dos casos, eles são divididos em: requisitos funcionais (estes representam o que o sistema deve possuir de funcionalidades na sua aplicação, ou seja, o que o sistema deve fazer) e não funcionais (estes representam os atributos do sistema, enquanto software constituído, incluindo manutenibilidade, eficiência, etc.). (NARDI e outros, 2006).

4.2.2.1 Requisitos Funcionais

Para Nardi e outros (2006), requisitos funcionais são uma subdivisão dos requisitos de software do sistema, estes denominam o que o sistema deve possuir de funcionalidades no seu funcionamento. O sistema deve atender a todos os requisitos especificados.

A seguir é apresentada a lista de requisitos funcionais do sistema proposto:

- RF001 – O sistema deve permitir que o usuário logado adicione documentos à plataforma;
- RF002 – O sistema deve permitir que outro usuário visualize a lista de documentos adicionados;
- RF003 – O sistema deve permitir que todos os usuários visualizem indicadores gerados a partir das referências bibliográficas contidas na base de dados;
- RF004 – O sistema deve permitir que os usuários se registrem na plataforma;
- RF005 – O sistema deve permitir que os usuários editem suas informações pessoais;

- RF006 – O sistema deve permitir que os usuários acessem à plataforma realizando login;

- RF007 – O sistema deve gerar indicadores a partir de referências bibliográficas contidas no banco de dados;

- RF008 – O sistema deve extrair informações de referências bibliográficas dos documentos inseridos pelo usuário;

4.2.2.2 Requisitos Não Funcionais

Nardi e outros (2006) afirmam que nos requisitos não funcionais são apresentados os atributos do sistema, sem serem necessariamente funções do sistema no seu desenvolvimento, são apresentadas ferramentas que o sistema irá executar e utilizar no seu funcionamento, dentre outras informações.

A seguir, é apresentada a lista de requisitos não funcionais do sistema:

- RNF001 – O sistema deve garantir que o retorno de registro de uma busca apresentará somente dados relevantes àquela consulta;

- RNF002 – Obrigatoriedade do uso do banco de dados PostgreSQL;

- RNF003 – Banco de dados deve suportar mais de mil registros em uma tabela;

- RNF004 – O sistema deverá permitir documentos do tipo txt.

4.2.2.3 Regras de Negócio

Regras de negócio são componentes do sistema de informação que representam um conceito dentro do processo de definição de requisitos do sistema que devem ser vistas como uma declaração genérica sobre a organização. É uma categoria de requisitos do sistema que representam decisões sobre como executar o negócio. (DE PÁDUA, 2001).

A seguir, é apresentada a lista de regras de negócio do sistema:

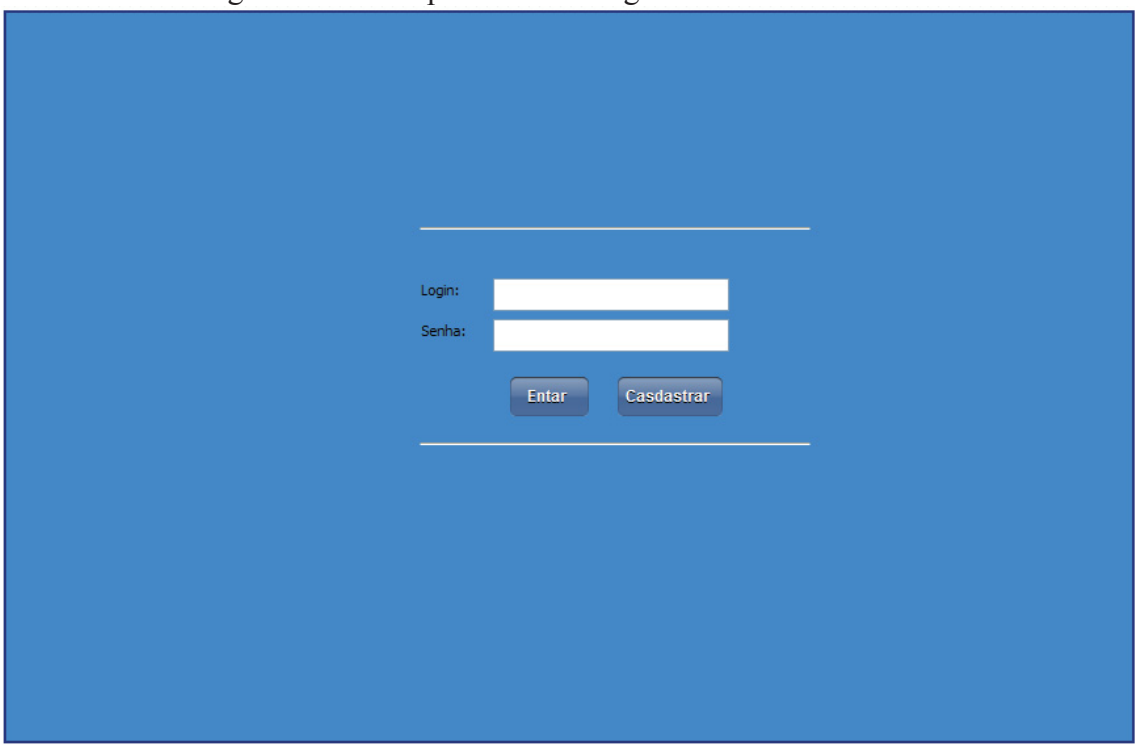
- RN001 – Para acessar o sistema deverá ser feito através do login cadastrado;
- RN002 – Somente usuários logados na plataforma poderão usar o sistema.

4.2.3 Protótipos de tela

Neste tópico, é apresentada uma ideia inicial, contendo os protótipos das telas do sistema desenvolvimento.

A figura 7 apresenta a tela inicial do sistema, chamada de tela de *login*, onde o usuário informa os dados de acesso cadastrados para usufruir das funcionalidades existentes.

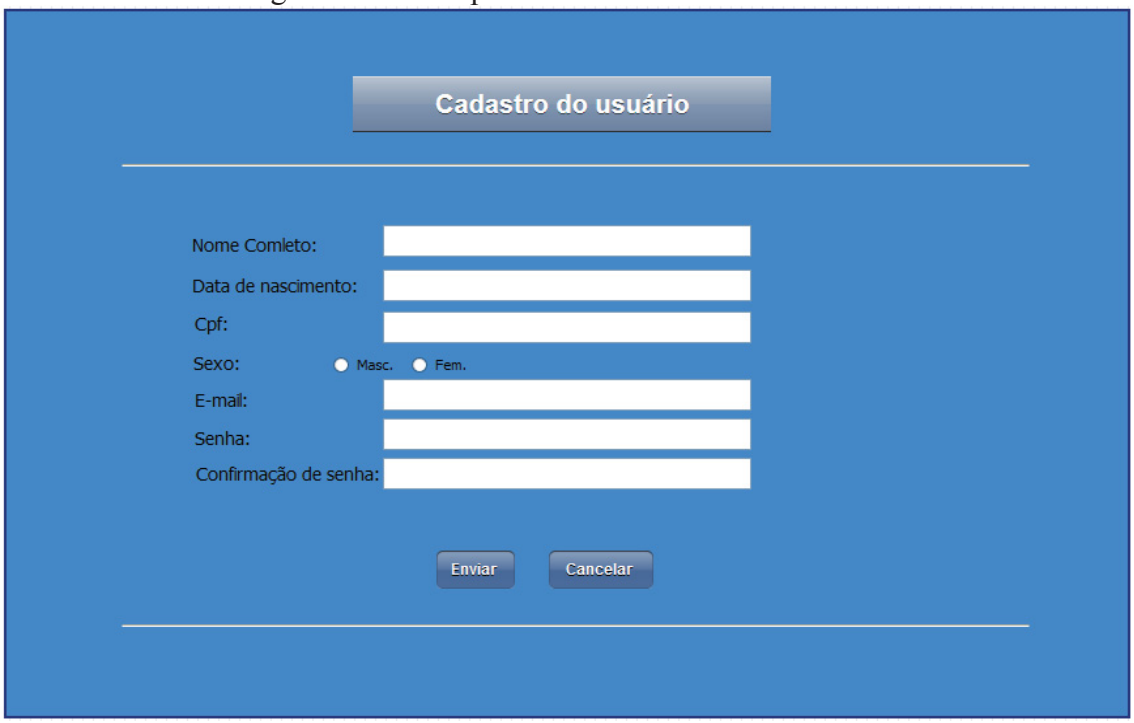
Figura 7 – Protótipo de tela de login de acesso ao sistema



Fonte: Autor, 2015.

A Figura 8 apresenta a tela de cadastro de usuário, etapa que é obrigatória para que o usuário realize seu cadastro no sistema e, com isso, consiga acessar suas funcionalidades.

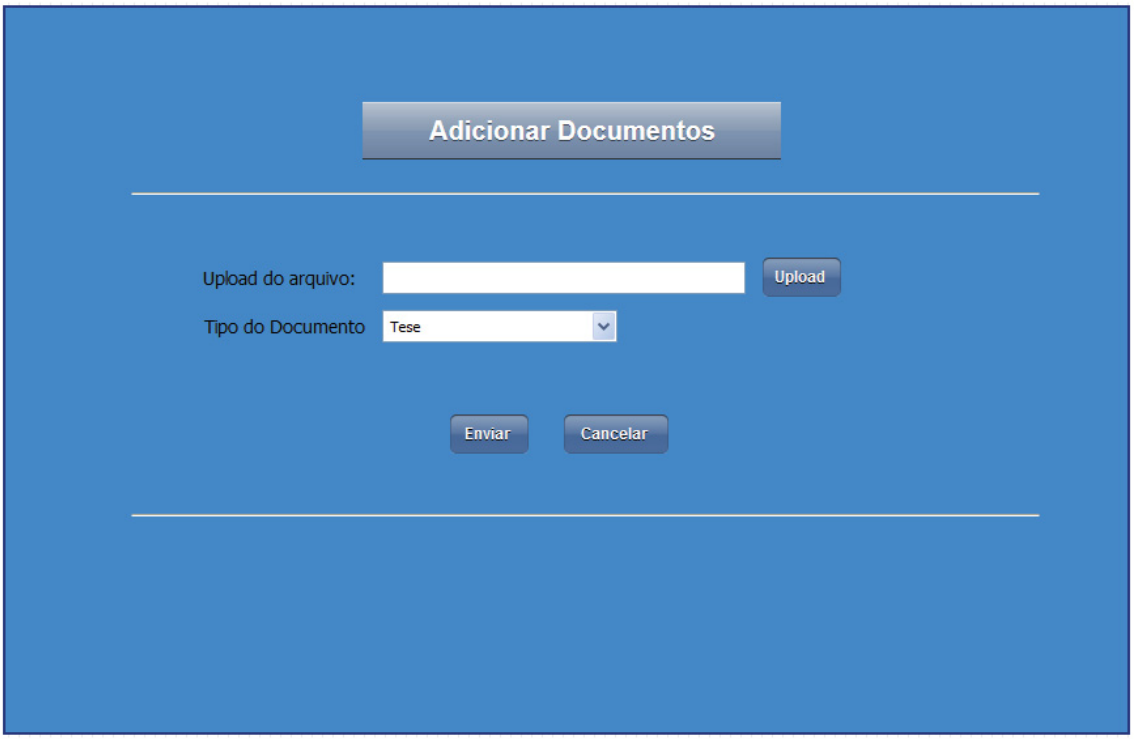
Figura 8 – Protótipo de tela de cadastro de usuário



Fonte: Autor, 2015.

A Figura 9 ilustra a tela de adição de documentos ao sistema, é através desta tela que é realizado a extração de dados do documento, para que possibilite, posteriormente, a geração de indicadores referentes aos dados inseridos. Ao adicionar um documento, o usuário deve informar a que tipo de documento este se refere, sendo possível visualizar a lista de documentos adicionados em outra tela do sistema, ilustrada na figura 10.

Figura 9 – Protótipo de tela de adição de documentos



Fonte: Autor, 2015.

A figura 10 ilustra a tela de lista de documentos adicionados pelo usuário, nela é possível visualizar o nome do documento, o tipo do documento e a data de inserção.

Figura 10 – Protótipo de tela de visualização de documentos adicionados

Fonte: Autor, 2015.

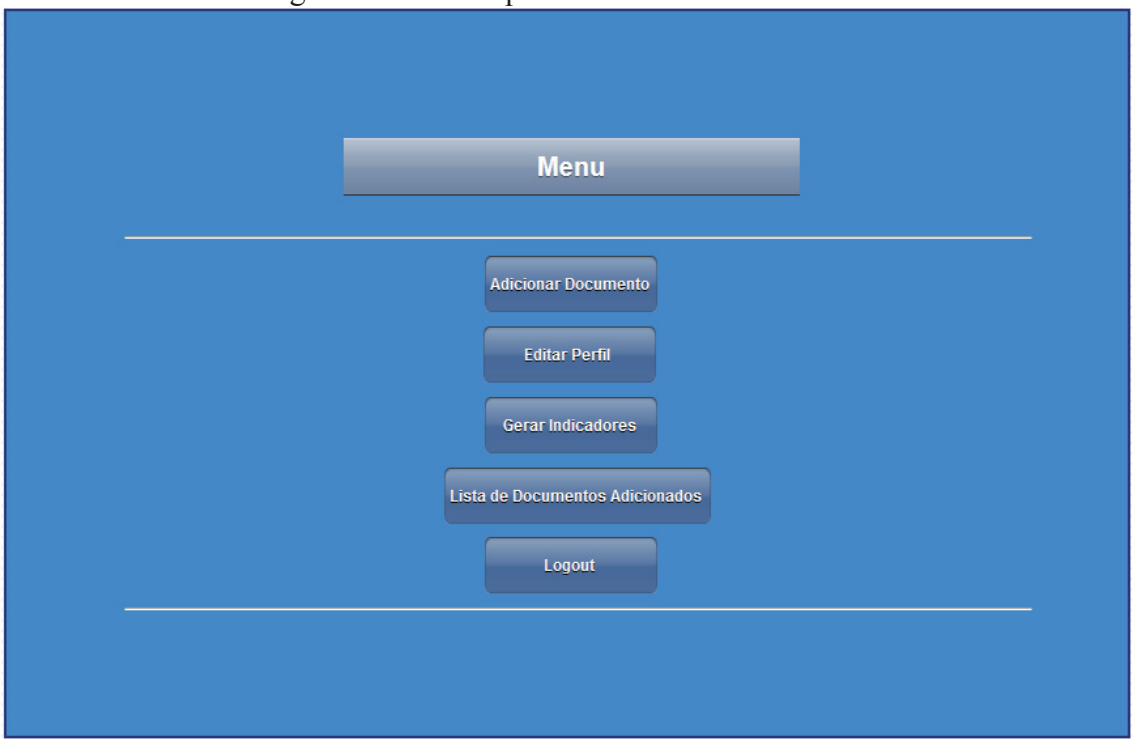
A figura 11 refere-se à tela de filtro para geração de indicadores, através dela, o usuário pode escolher as variáveis desejadas para geração dos dados.

Figura 11 – Protótipo de tela de filtro para a geração de indicadores.

Fonte: Autor, 2015.

Na figura 12 é ilustrada a tela inicial do sistema, acessada após o usuário realizar o *login* do sistema, a tela possibilita direcionar o usuário para as funcionalidades desejadas.

Figura 12 – Protótipo da tela inicial do sistema.



Fonte: Autor, 2015

Neste tópico, foi possível ter uma noção inicial do sistema, visualizando os protótipos ilustrados por figuras, apresentando as funcionalidades do sistema. O próximo tópico dá andamento à modelagem do sistema, nele são apresentados os casos do sistema e sua modelagem.

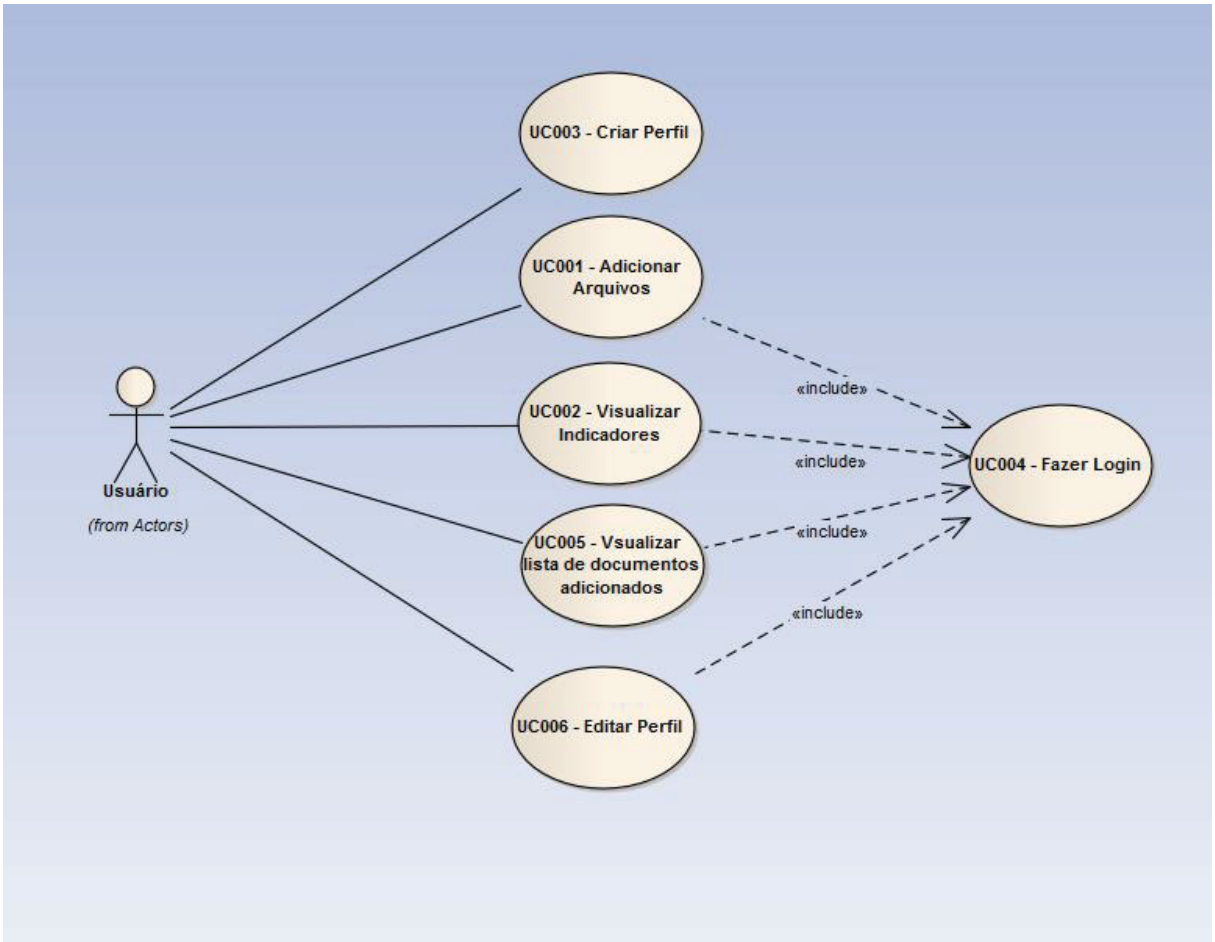
4.2.4 Casos de Uso

Nesta seção, é apresentado o diagrama de caso de uso, seguindo a metodologia de desenvolvimento IConix. Cada caso de uso é descrito, contendo seus fluxos principais, alternativos e de exceção.

Este modelo de caso de uso é usado para representar as exigências do usuário seja um sistema novo (partindo do nada) ou baseado em um sistema já existente. O modelo de caso de uso tem o objetivo de detalhar de forma clara e legível, todos os cenários que os usuários executarão para realizar alguma tarefa. (MAIA, 2005).

A figura 13 demonstra os casos de uso que são apresentados no sistema, em seguida o detalhamento de cada um destes casos de uso.

Figura 13 – Representação dos casos de uso do sistema



Fonte: Autor, 2015.

UC004 – Fazer Login (Fluxo Principal)

1. O usuário acessa a plataforma
2. O sistema apresenta a tela de login
3. O usuário informa os dados de acesso (e-mail e senha)

4. O sistema valida o dados informados pelo usuário
5. O sistema apresenta a tela inicial do sistema
6. Fim do fluxo principal

UC004 – Fazer Login (Fluxo de Exceção)

1. O usuário informa os dados de acesso incorretos
2. O sistema verifica que os dados que usuário informou não são reais
3. O sistema apresenta uma mensagem informando a falha no acesso
4. Fim do fluxo de exceção

UC003 – Criar Perfil (Fluxo Principal)

1. O usuário acessa a plataforma
2. O sistema apresenta a tela de login
3. O usuário informa clica no botão para registrar-se a plataforma
4. O sistema apresenta a tela de cadastro de usuário
5. O usuário pode cancelar a operação a qualquer momento
6. O usuário informa os dados para o cadastro
7. O sistema valida os dados informados pelo usuário
8. O sistema registra novo usuário na plataforma
9. O sistema apresenta a tela inicial do sistema
10. Fim do fluxo principal

UC003 – Criar Perfil (Fluxo de Exceção)

1. O usuário informa os dados de cadastro incorretos
2. O sistema informa quais campos não estão corretos
3. Fim do fluxo de exceção

UC003 – Criar Perfil (Fluxo de Exceção)

1. O usuário informa um e-mail já existente na plataforma

2. O sistema informa que e-mail já existe
3. Fim do fluxo de exceção

UC001 – Adicionar Arquivos (Fluxo Principal)

1. O usuário seleciona a opção para adicionar arquivos
2. O sistema apresenta a tela de adição de arquivos
3. O usuário adiciona um arquivo através do upload
4. O usuário informa o tipo de arquivo
5. O usuário clica em enviar
6. O sistema valida as informações
7. O sistema registra o documento na plataforma
8. Fim do fluxo principal

UC001 – Adicionar Arquivos (Fluxo de Exceção)

1. O usuário adiciona um arquivo através do upload
2. O usuário informa o tipo de arquivo
3. O usuário clica em enviar
4. O sistema verifica que os dados informados não são validos
5. O sistema apresenta uma mensagem de erro ao usuário
6. Fim do fluxo de exceção.

UC005 – Visualizar lista de documentos adicionados (Fluxo de Principal)

1. O usuário seleciona a opção para visualizar lista de documentos
2. O sistema apresenta os documentos adicionados pelo usuário
3. Fim do fluxo principal

UC005 – Visualizar lista de documentos adicionados (Fluxo Alternativo)

1. O usuário seleciona a opção para visualizar lista de documentos
2. O sistema verifica que o usuário não adicionou nenhum arquivo
3. O sistema a lista vazia com mensagem sugerindo ao usuário adicionar um documento.
4. Fim do fluxo alternativo

UC006 – Editar perfil (Fluxo Principal)

1. O usuário seleciona a opção para visualizar perfil
2. O sistema apresenta os dados pessoais do usuário
3. O usuário informa os novos dados
4. O sistema valida os dados de acesso do usuário
5. Fim do fluxo principal

UC006 – Editar perfil (Fluxo de Exceção)

1. O sistema valida os dados informados pelo usuário
2. O sistema verifica que os dados informados não estão corretos
3. O sistema apresenta uma mensagem de erro ao usuário
4. Fim do fluxo de exceção

UC002 – Visualizar indicadores (Fluxo Principal)

1. O usuário seleciona a opção para visualizar indicadores
2. O sistema apresenta a tela de filtros para a geração de indicadores
3. O informa os campos para filtro
4. O sistema valida as informações adicionadas pelo usuário
5. O sistema realiza a geração de indicadores
6. O sistema apresenta os indicadores através de gráficos
7. Fim do fluxo principal

UC002 – Visualizar indicadores (Fluxo Alternativo)

1. O usuário seleciona a opção para visualizar indicadores
2. O sistema apresenta a tela de filtros para a geração de indicadores
3. O informa os campos para filtro
4. O sistema valida as informações adicionadas pelo usuário
5. Verifica que não há registro com os filtros informados
6. O sistema apresenta uma mensagem ao usuário
7. Fim do fluxo alternativo

UC002 – Visualizar indicadores (Fluxo de Exceção)

1. O sistema verifica que os dados informados estão incorretos
2. O sistema apresenta uma mensagem ao usuário
3. Fim do fluxo de exceção

Neste tópico foram apresentados todos os cenários existentes dentro do funcionamento do sistema, o próximo tópico apresenta outra parte do modelo ICONIX, chamado de modelo do domínio.

4.2.5 Modelo de Domínio

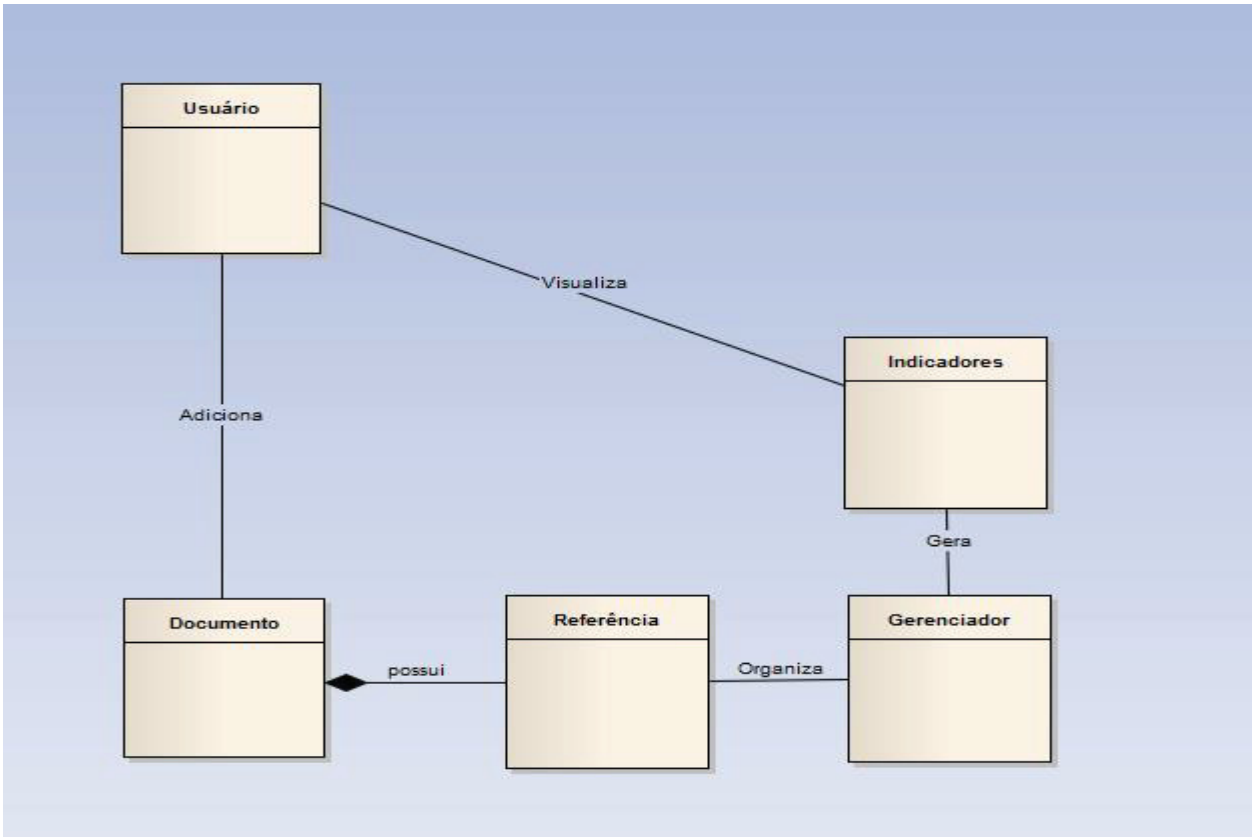
Segundo Maia (2005, p.4), "O Modelo de Domínio é uma parte essencial do processo de ICONIX. Ele constrói uma porção estática inicial de um modelo que é essencial para dirigir a fase de design a partir dos casos de uso.".

Para realizar o modelo de domínio, é preciso tentar descobrir o maior número possível de classes existentes no problema para qual se pretende desenvolver o software. Claro que modelo de domínio não retratará o cenário completo adequando para o problema

que se pretende resolver, algumas classes serão excluídas e outras serão encontradas ou modificadas etc. Isso faz parte do processo de desenvolvimento de softwares.

Na figura 14, é apresentado o modelo de domínio. Através dele, é possível ter uma noção das principais classes do sistema.

Figura 14 – Modelo de Domínio do Sistema



Fonte: Autor, 2015.

O principal objetivo do Modelo de Domínio é descobrir objetos e associações, portanto, detalhes de cada objeto e sua multiplicidade devem ser omitidos.

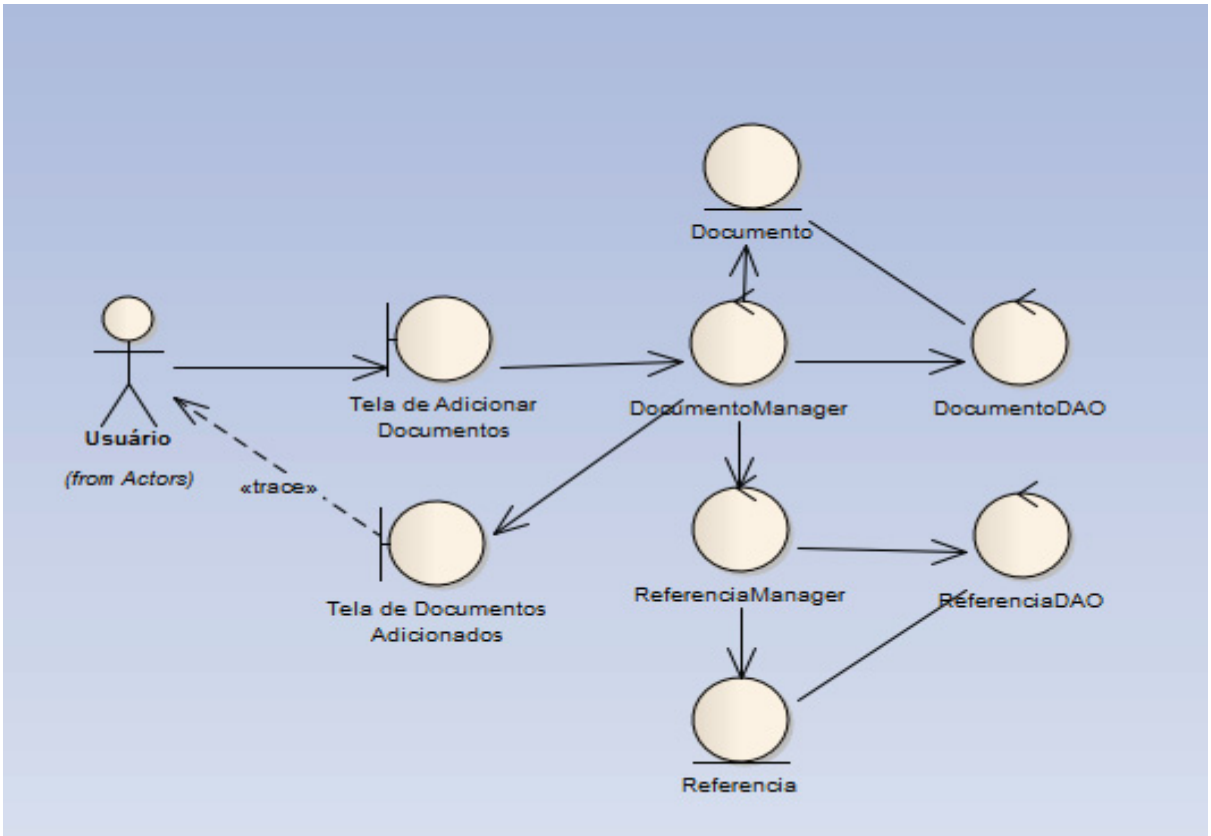
Neste tópico, foram apresentadas as principais classes que compõem o modelo de domínio do sistema, no próximo tópico, é apresentado o modelo de robustez, também chamado de Análise de Robustez ou Análise Robusta.

4.2.6 Análise de Robustez

Nesta seção, é apresentado o modelo de robustez. Nele podemos notar o fluxo principal de execução de cada caso de uso que compõe o sistema. Após cada modelo, é relatada uma breve descrição do funcionamento do modelo.

"A Análise Robusta focaliza construir um modelo, analisando as narrativas de texto de caso de uso, identificando um conjunto de objetos que participarão de cada caso de uso". (MAIA, 2005, p.8).

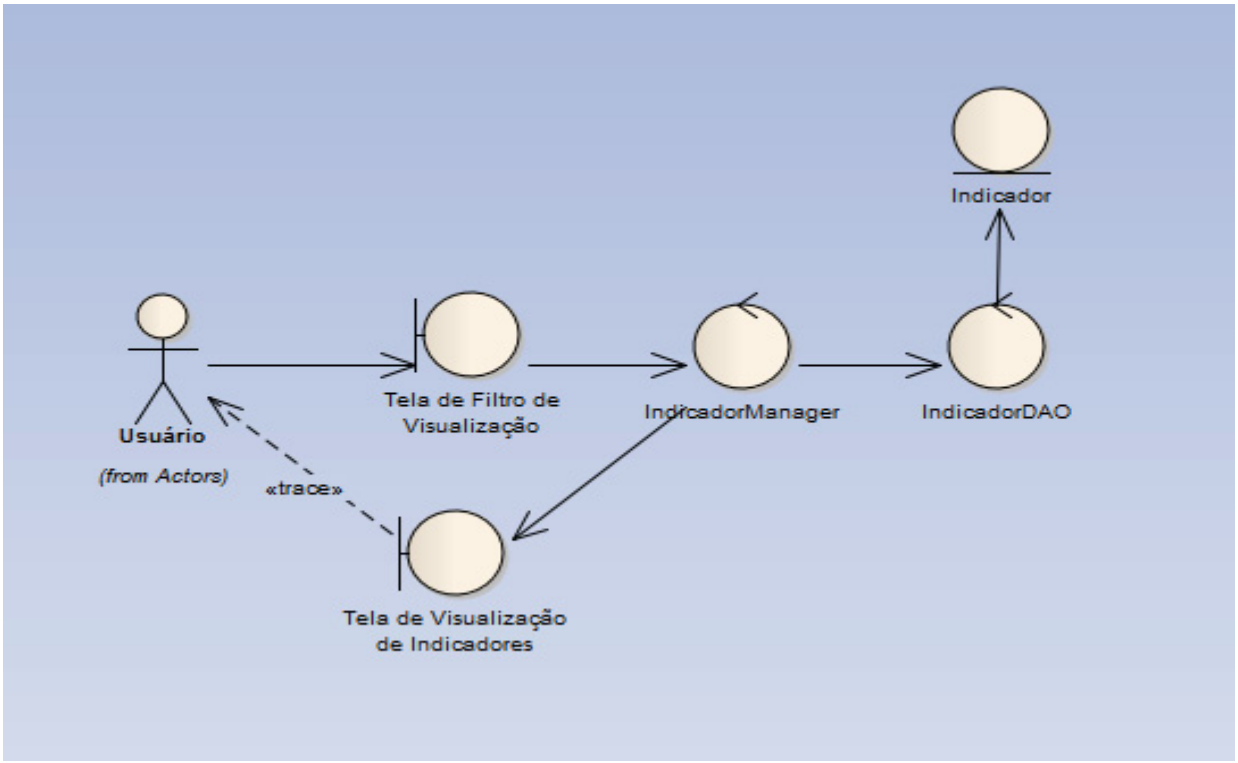
Figura 15 – Modelo de Robustez do UC001 (Adicionar Documentos)



Fonte: Autor, 2015.

Na figura 15, é apresentado o fluxo principal de funcionamento do caso de uso de adicionar documentos à plataforma. Nele o usuário, através da tela de adição de documentos, realiza a inserção com os devidos dados. Então, o sistema valida os dados de acesso através da classe DocumentoManager e realiza a extração de dados, retirando as referências bibliográficas posteriormente na classe ReferenciaManager. Em seguida, é retirada cada informação das referências bibliográficas (autor, ano de publicação, etc.) e, então, essas informações são armazenadas no banco de dados.

Figura 16 – Modelo de Robustez do UC002 (Visualizar Indicadores)

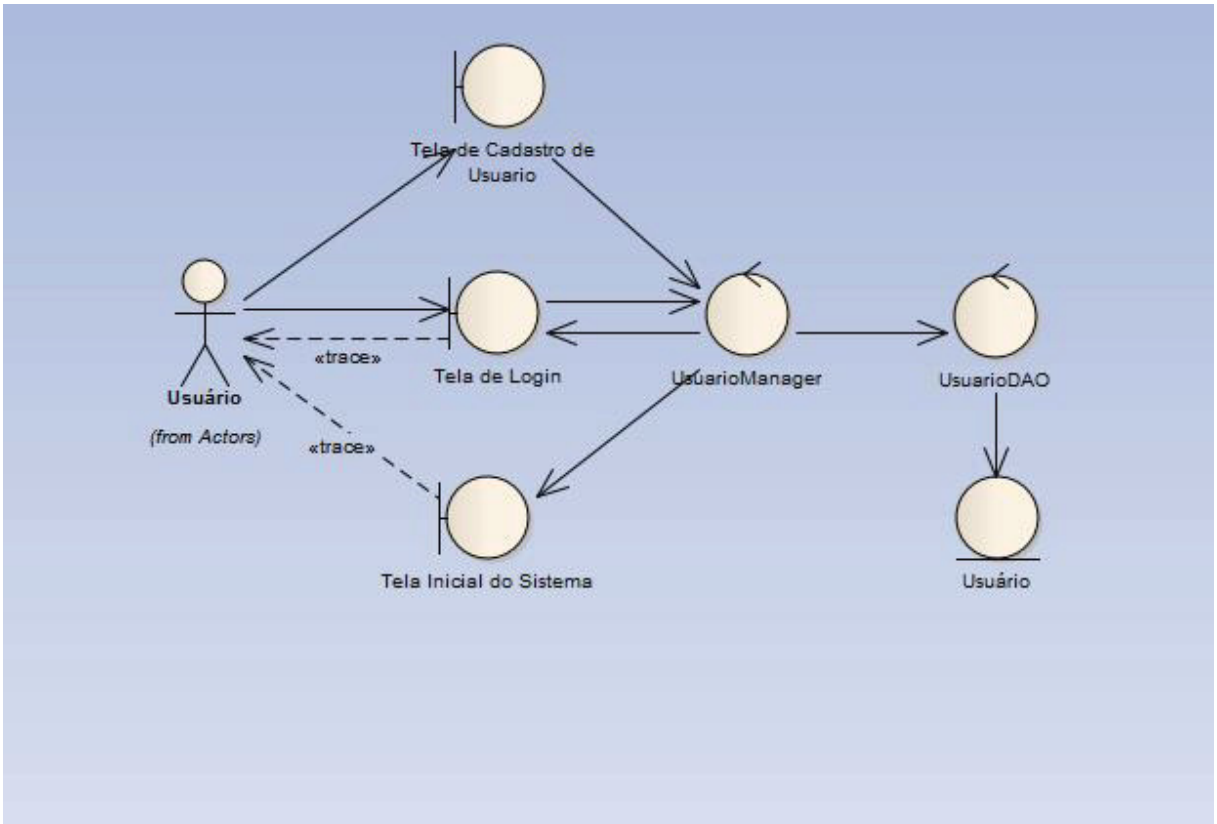


Fonte: Autor, 2015.

Na figura 16, é apresentado o fluxo principal do caso de uso de visualizar indicadores. Nele, o usuário informa os filtros para a geração de indicadores, em seguida, o sistema valida os campos e realiza a busca pelas informações filtradas.

Após realizar a busca, o sistema manipula as informações para a geração de indicadores e apresenta, através de gráficos, o resultado gerado.

Figura 17 – Modelo de Robustez do UC003 (Criar Perfil)

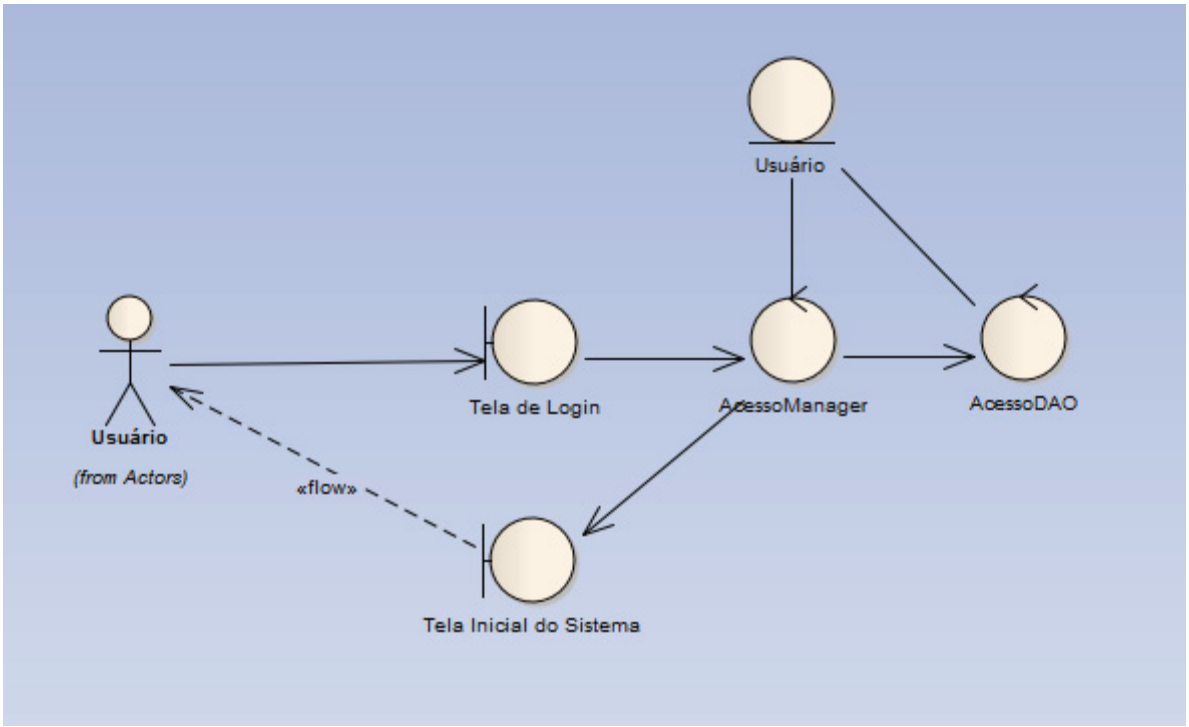


Fonte: Autor, 2015.

Na figura 17 é apresentado o fluxo principal do caso de uso de criar perfil. Nele, o usuário, através da tela de cadastro, informa seus dados pessoais, em seguida, o sistema realiza a validação dos dados e registra o mesmo na plataforma.

Após o registro, o sistema encaminha o usuário para a tela inicial do sistema.

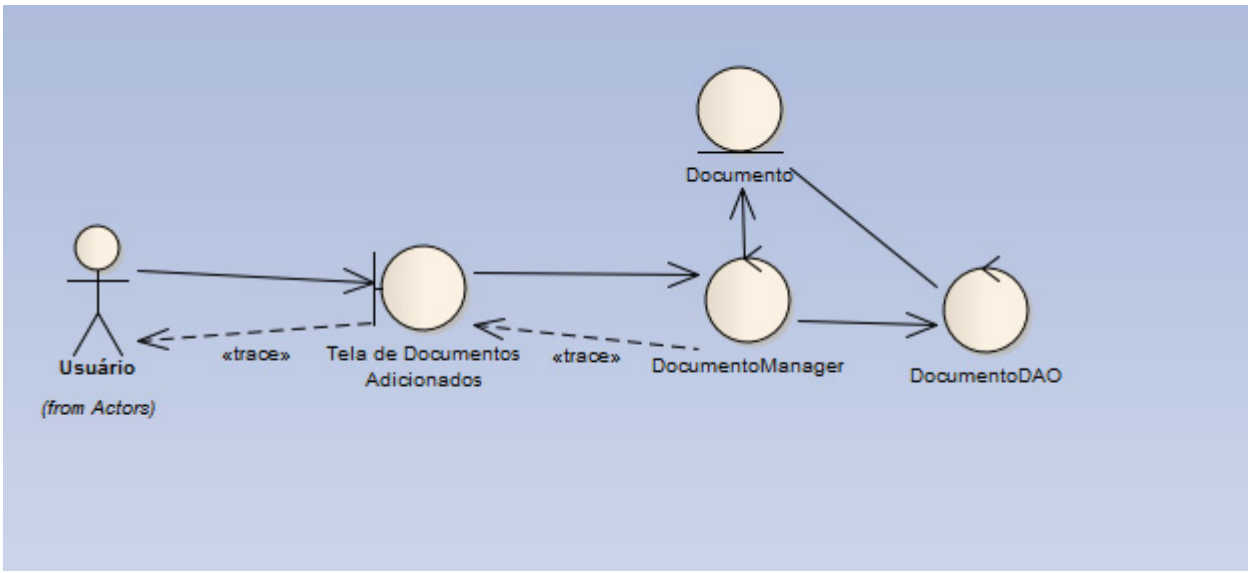
Figura 18 – Modelo de Robustez do UC004 (Fazer Login)



Fonte: Autor, 2015.

Na figura 18, é apresentado o fluxo principal do caso de uso de realizar login. O usuário informa os dados de acesso à plataforma, em seguida, as informações são validadas e, caso estejam de acordo, o sistema encaminha o usuário para a tela principal do sistema.

Figura 19 – Modelo de Robustez do UC005 (Visualizar Lista de Documentos Adicionados)

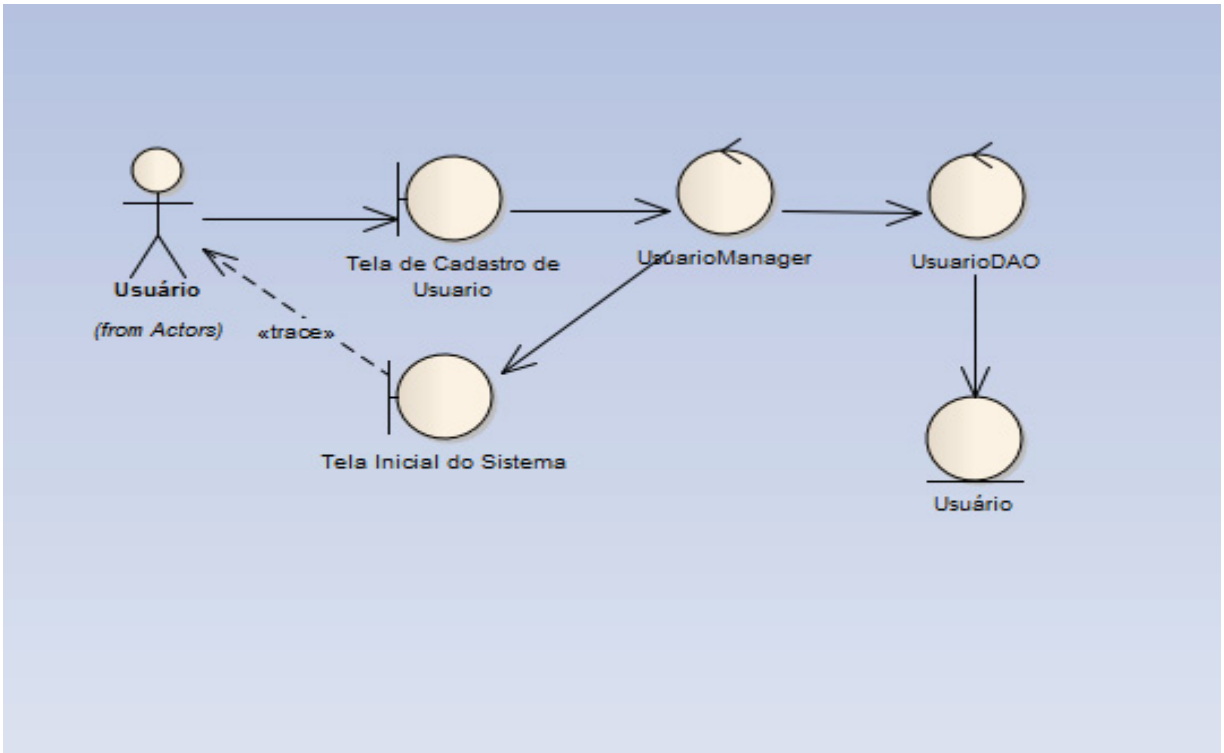


Fonte: Autor, 2015.

Na Figura 19, é apresentado o fluxo principal do caso de uso de visualizar lista de documentos adicionados. Nela, o usuário seleciona a opção no menu principal para visualizar

a lista de documentos. O sistema realiza a busca dos documentos do usuário no banco e retorna a lista ao usuário.

Figura 20 – Modelo de Robustez do UC006 (Editar Perfil)



Fonte: Autor, 2015.

Na figura 20, é apresentado o fluxo principal do caso de uso de editar perfil. Nele, o usuário, através da tela de cadastro, preenche as informações editadas. O sistema valida as informações e registra os dados do usuário na plataforma.

Nesta seção, pode-se visualizar o modelo de robustez que compõe a modelagem do sistema proposto. No próximo tópico, é apresentado o modelo de sequência que ilustra este fluxo de forma detalhada.

4.2.7 Diagrama de Sequencia

Nesta seção, é apresentado o modelo de sequência. Nele, podemos notar o fluxo principal de execução detalhado de cada caso de uso que compõe o sistema. Após cada modelo, é relatada uma breve descrição do funcionamento do modelo.

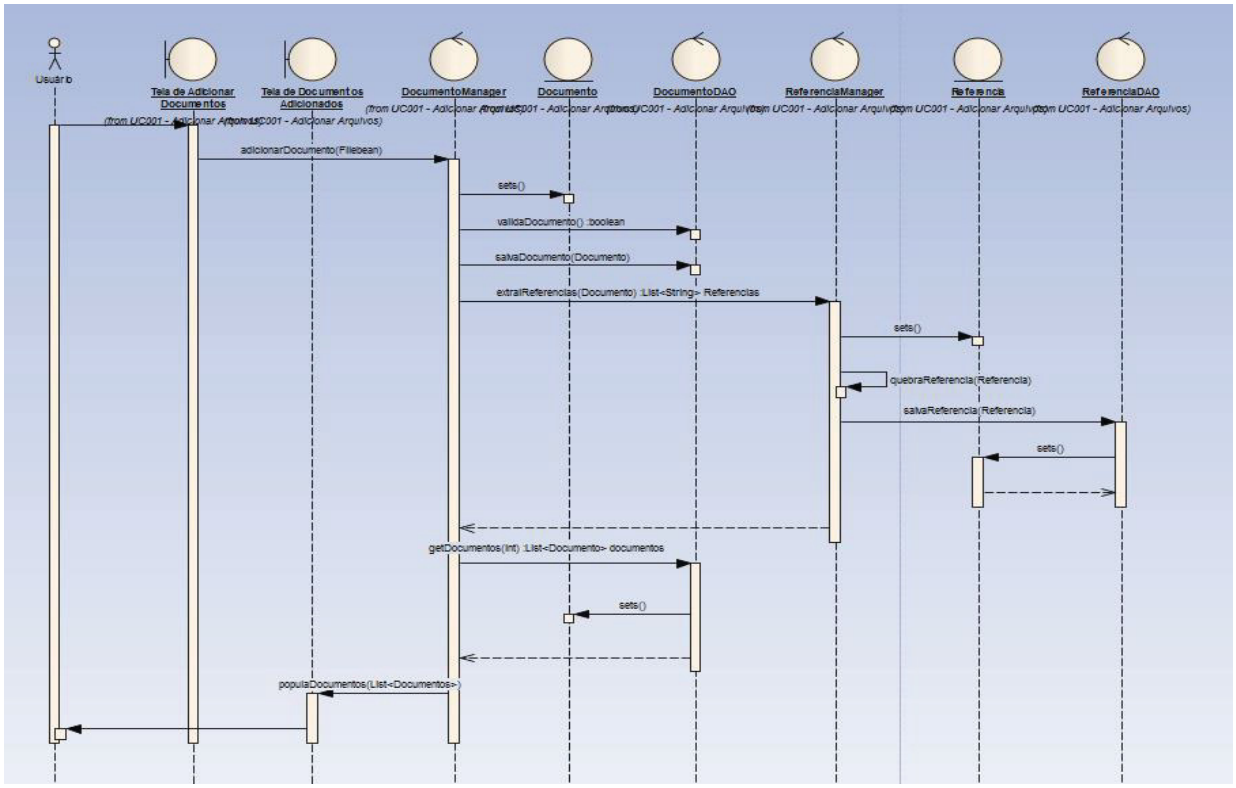
> "O Diagrama de Seqüência tem como objetivo construir um modelo dinâmico entre o usuário e o sistema. Para tal, devemos utilizar os objetos e suas interações identificadas na análise robusta, só que agora, temos por obrigação o detalhamento de cada fluxo de ação. Teoricamente isso já é possível, pois, uma vez concluído o modelo de domínio (diagrama de classe de alto nível) e análise robusta, nós, teremos descoberto a maioria dos objetos no contexto do problema e definido alguns atributos e relações estáticas entre objetos no diagrama de classe de alto nível e algumas relações dinâmicas na análise robusta. Estas conquistas representam passos largos para um bom resultado.". (MAIA, 2005, p.12).

No modelo de sequência, é demonstrado o cenário mais detalhado que antecede o desenvolvimento, nele, são apresentados desde os métodos que serão utilizados para a execução da tarefa, até os atributos necessários.

A seguir, é descrito o modelo de sequência para cada um dos Casos de Uso descritos no sistema:

A figura 21 descreve o modelo de sequência para o Caso de Uso 001 – Adicionar documentos no sistema, nele, é apresentado todo o fluxo executado para que essa funcionalidade possa realizar a ação necessária esperada do sistema.

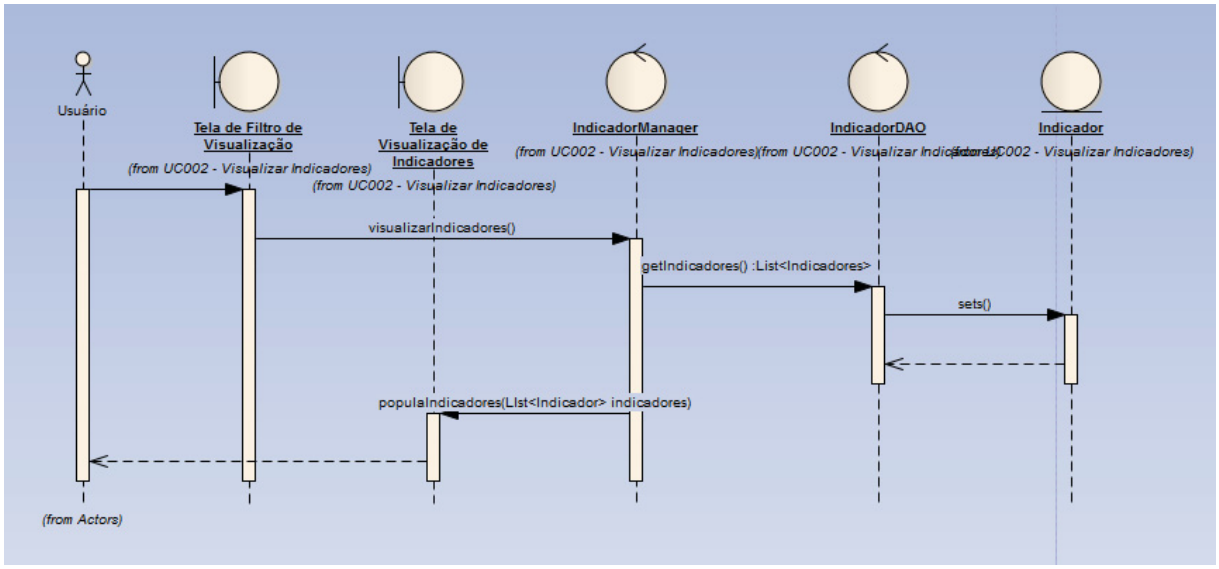
Figura 21 – Modelo de Sequência do UC001 (Adicionar Documentos ao Sistema)



Fonte: Autor, 2015.

A figura 22 descreve o modelo de sequência para o Caso de Uso 002 – Visualizar Indicadores, nele é apresentado todo o fluxo executado para que esta funcionalidade possa realizar a ação necessária esperada do sistema.

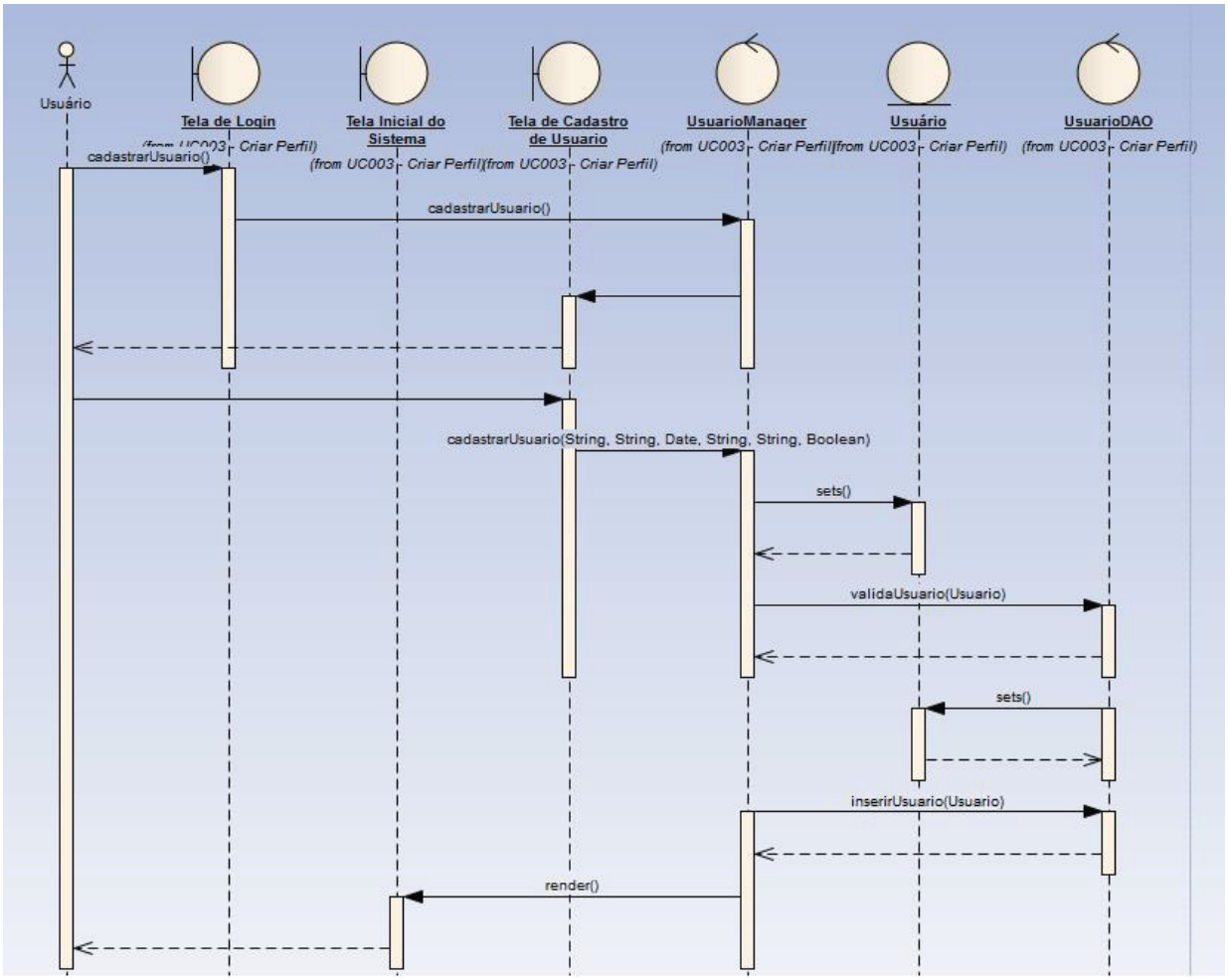
Figura 22 – Modelo de Sequência do UC002 (Visualizar indicadores)



Fonte: Autor, 2015.

A figura 23 descreve o modelo de sequência para o Caso de Uso 003 – Criar Perfil, nele, é apresentado todo o fluxo executado para que esta funcionalidade possa realizar a ação necessária esperada do sistema.

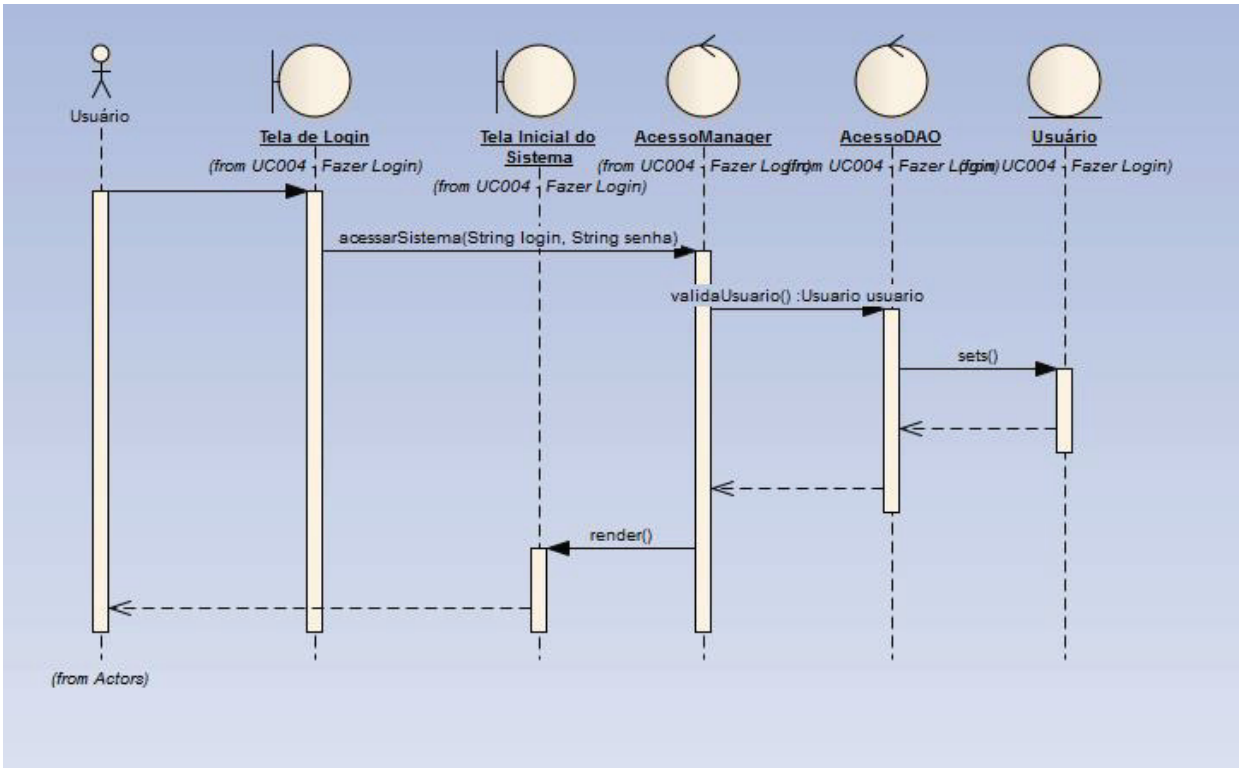
Figura 23 – Modelo de Sequência do UC003 (Criar Perfil)



Fonte: Autor, 2015.

A figura 24 descreve o modelo de sequência para o Caso de Uso 004 – Fazer Login, nele é apresentado todo o fluxo executado para que esta funcionalidade possa realizar a ação necessária esperada do sistema.

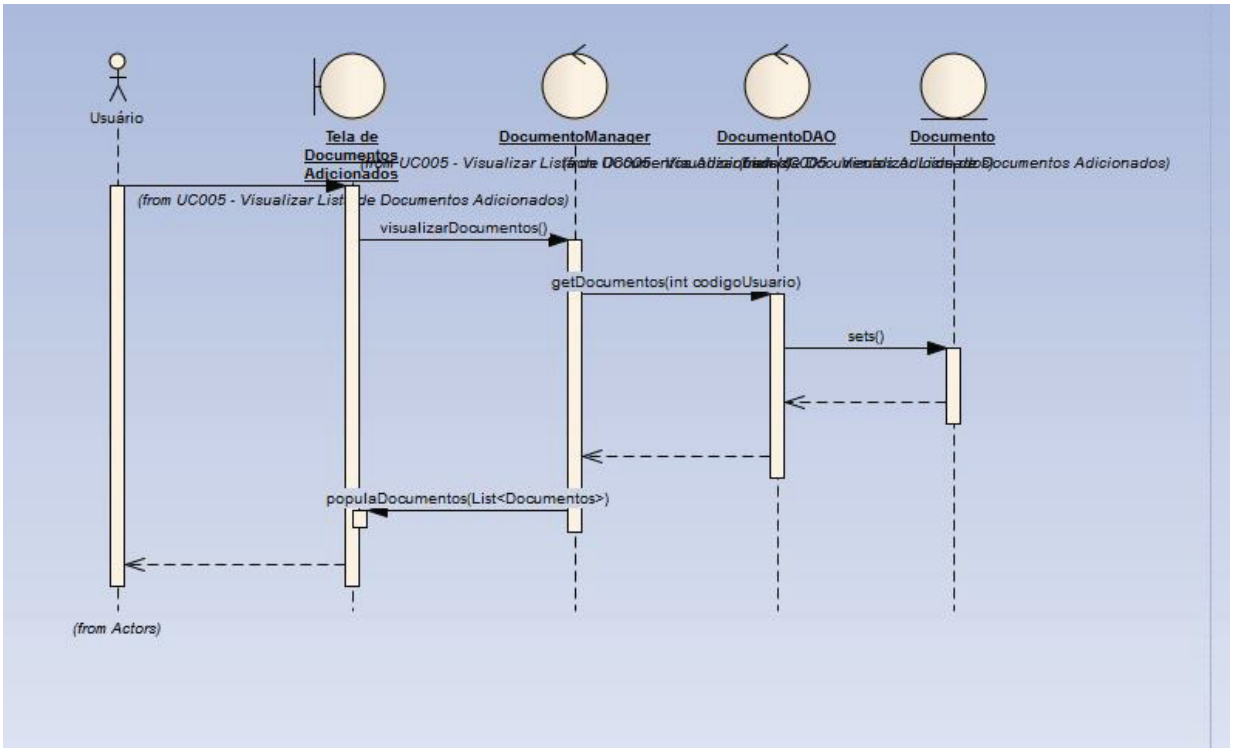
Figura 24 – Modelo de Sequência do UC004 (Fazer Login)



Fonte: Autor, 2015.

A figura 25 descreve o modelo de sequência para o Caso de Uso 005 – Visualizar Lista de Documentos Adicionados, nele é apresentado todo o fluxo executado para que esta funcionalidade possa realizar a ação necessária esperada do sistema.

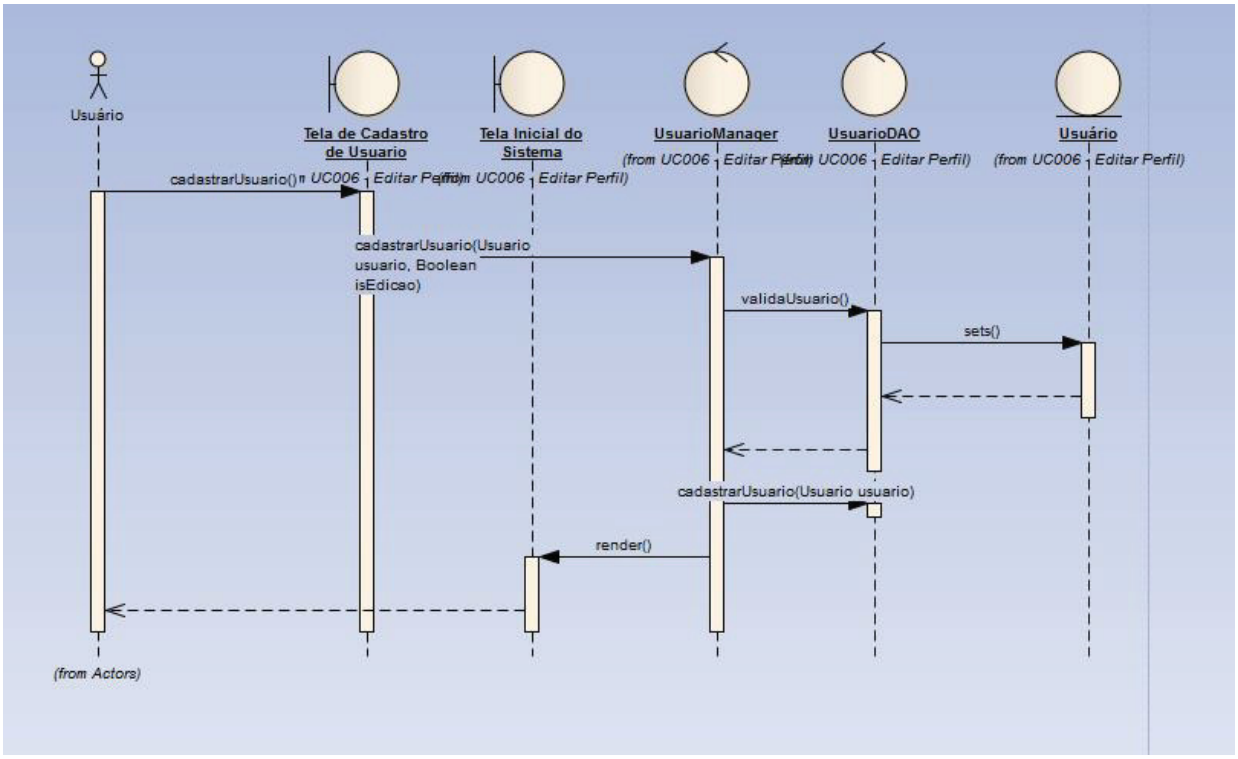
Figura 25 – Modelo de Sequência do UC005 (Visualizar Lista de Documentos Adicionados)



Fonte: Autor, 2015.

A figura 26 descreve o modelo de sequência para o Caso de Uso 005 – Visualizar Lista de Documentos Adicionados, nele é apresentado todo o fluxo executado para que esta funcionalidade possa realizar a ação necessária esperada do sistema.

Figura 26 – Modelo de Sequência do UC006 (Editar Perfil)



Fonte: Autor, 2015.

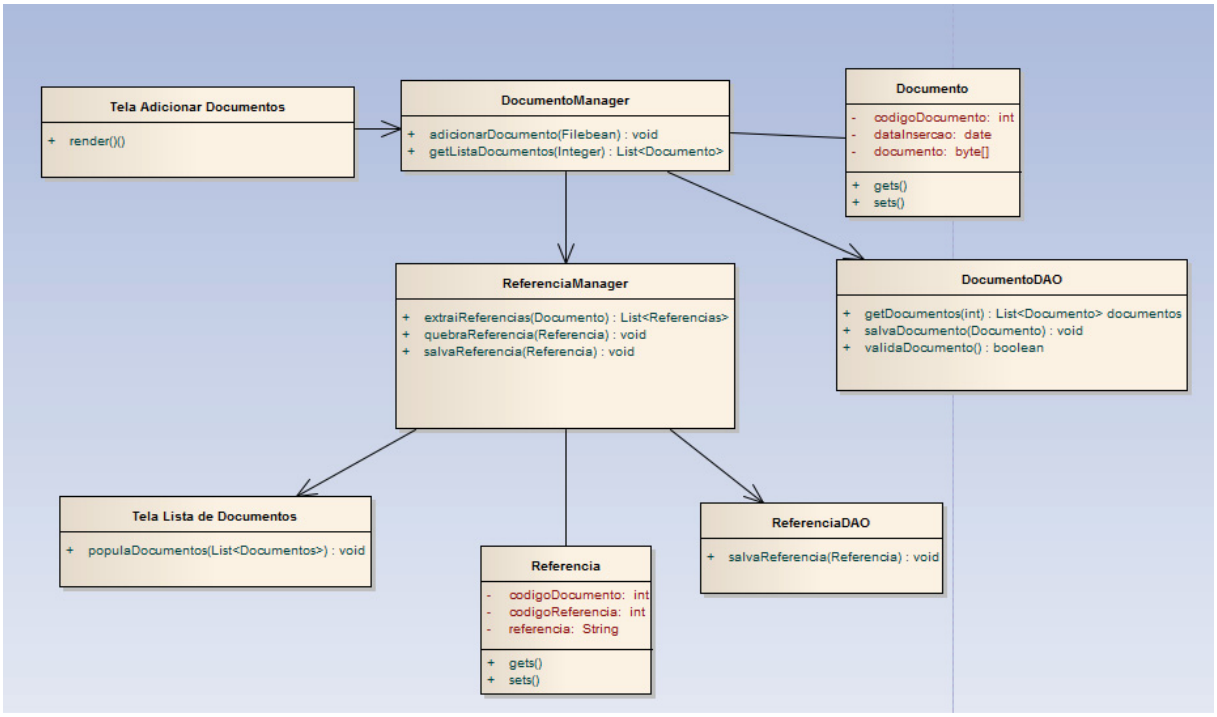
Nesta seção, foi apresentado o modelo de sequência que compõe a modelagem do sistema proposto. No próximo tópico, é apresentado o modelo de classe que ilustra as classes do sistema.

4.2.8 Diagrama de Classes

O diagrama de classe é um dos modelos que compõe o desenvolvimento utilizando o ICONIX, ele é desenvolvido a partir do princípio do modelo de domínio e representa as funcionalidades do sistema de modo estático sem a interação do usuário com o sistema. (MAIA, 2005).

A Figura 27 demonstra o Modelo de Classe necessário para o desenvolvimento do Caso de Uso 001 – Adicionar Documentos no Sistema.

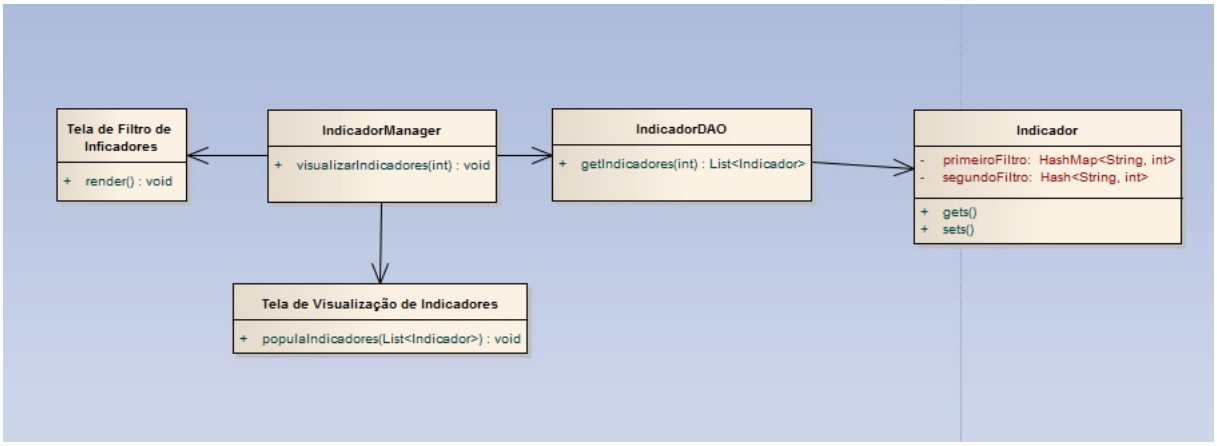
Figura 27 – Modelo de Classe do UC001 (Adicionar Documentos no Sistema)



Fonte: Autor, 2015.

A Figura 28 demonstra o Modelo de Classe necessário para o desenvolvimento do Caso de Uso 002 – Visualizar Indicadores.

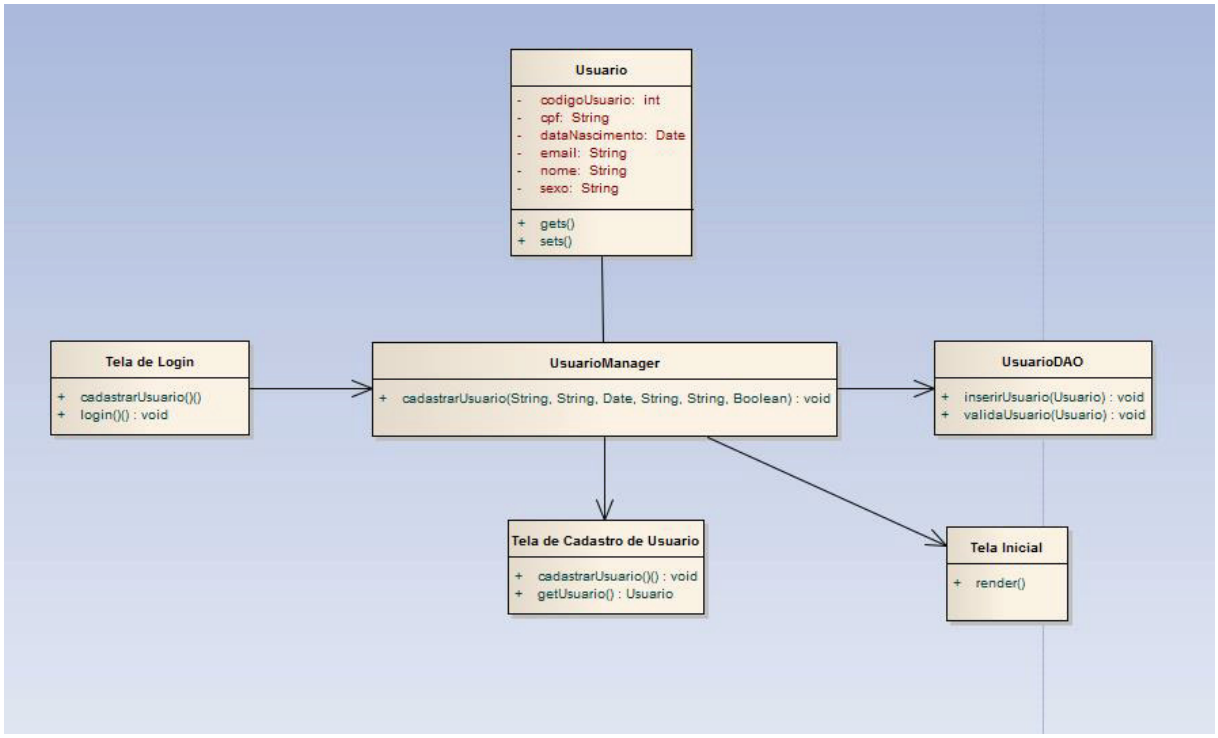
Figura 28 – Modelo de Classe do UC002 (Visualizar Indicadores)



Fonte: Autor, 2015.

A Figura 29 demonstra o Modelo de Classe necessário para o desenvolvimento do Caso de Uso 003 – Criar Perfil.

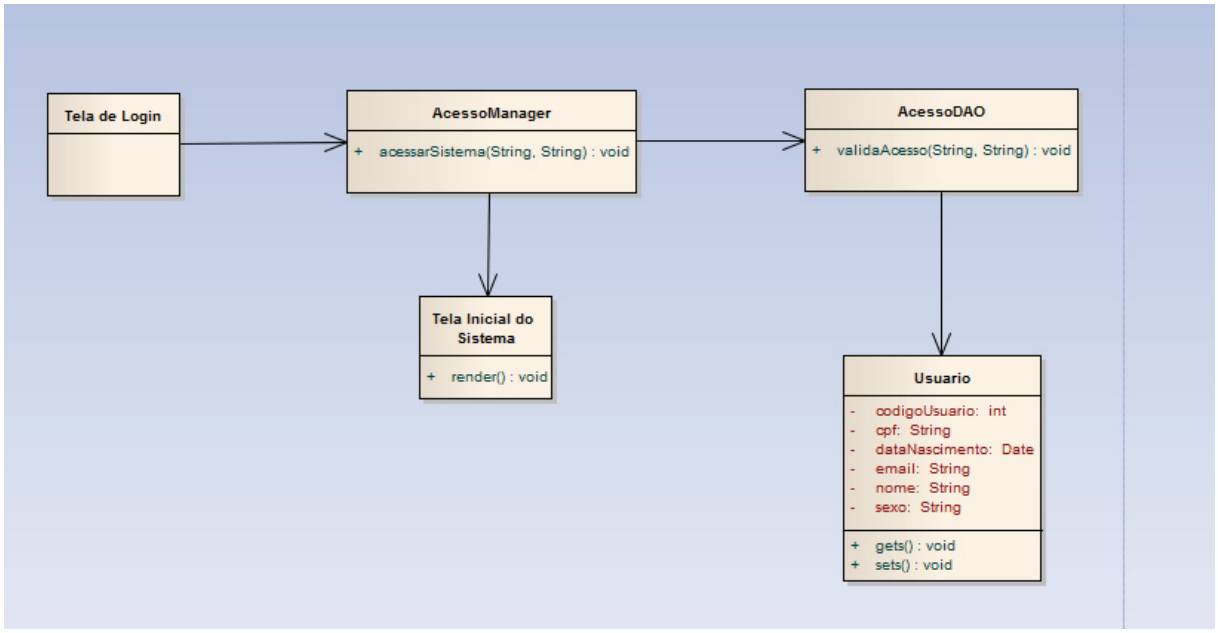
Figura 29 – Modelo de Classe do UC003 (Criar Perfil)



Fonte: Autor, 2015.

A Figura 30 demonstra o Modelo de Classe necessário para o desenvolvimento do Caso de Uso 004 – Fazer Login.

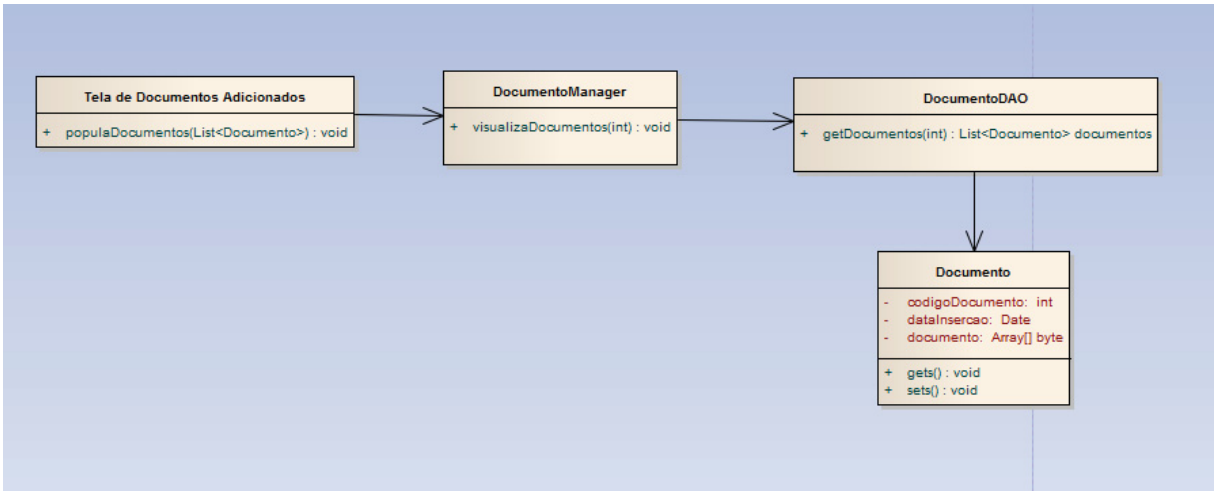
Figura 30 – Modelo de Classe do UC004 (Fazer Login)



Fonte: Autor, 2015.

A Figura 31 demonstra o Modelo de Classe necessário para o desenvolvimento do Caso de Uso 005 – Visualizar Lista de Documentos Adicionados.

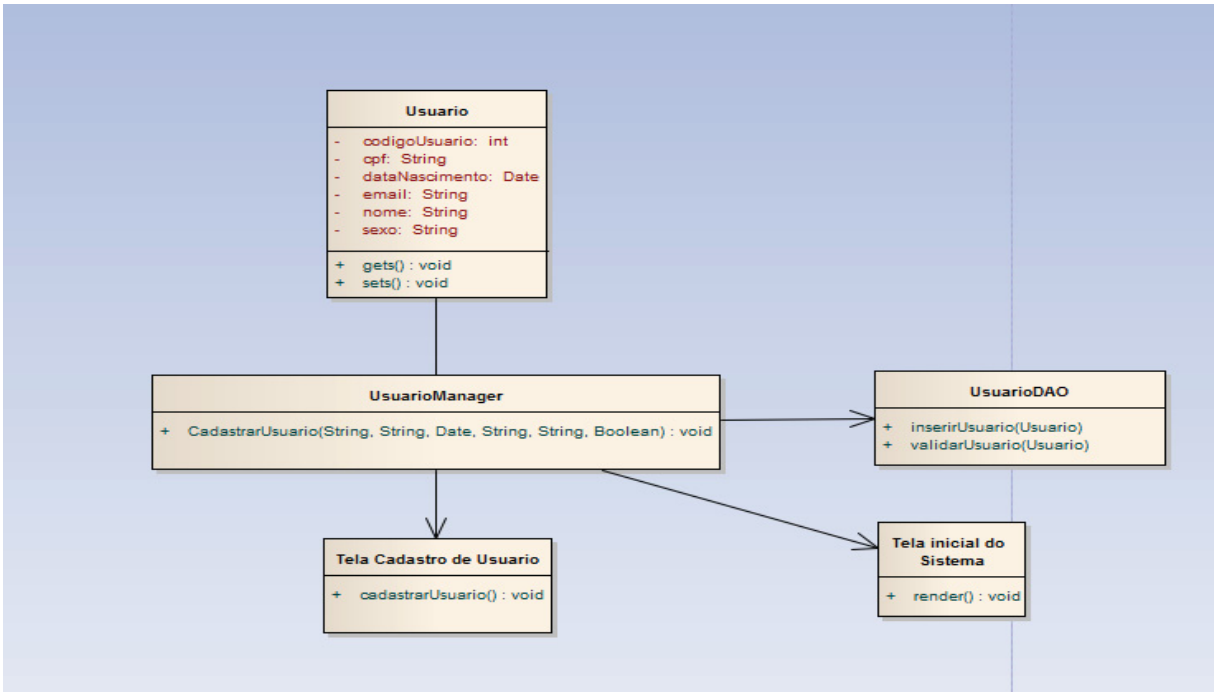
Figura 31 – Modelo de Classe do UC005 (Visualizar Lista de Documentos Adicionados)



Fonte: Autor, 2015.

A Figura 32 demonstra o Modelo de Classe necessário para o desenvolvimento do Caso de Uso 006 – Editar Perfil.

Figura 32 – Modelo de Classe do UC006 (Editar Perfil)



Fonte: Autor, 2015.

Nesta seção, pôde-se entender um pouco mais sobre metodologia de desenvolvimento ICONIX e visualizar todos os cenários existentes em todos os modelos que compõem esta metodologia.

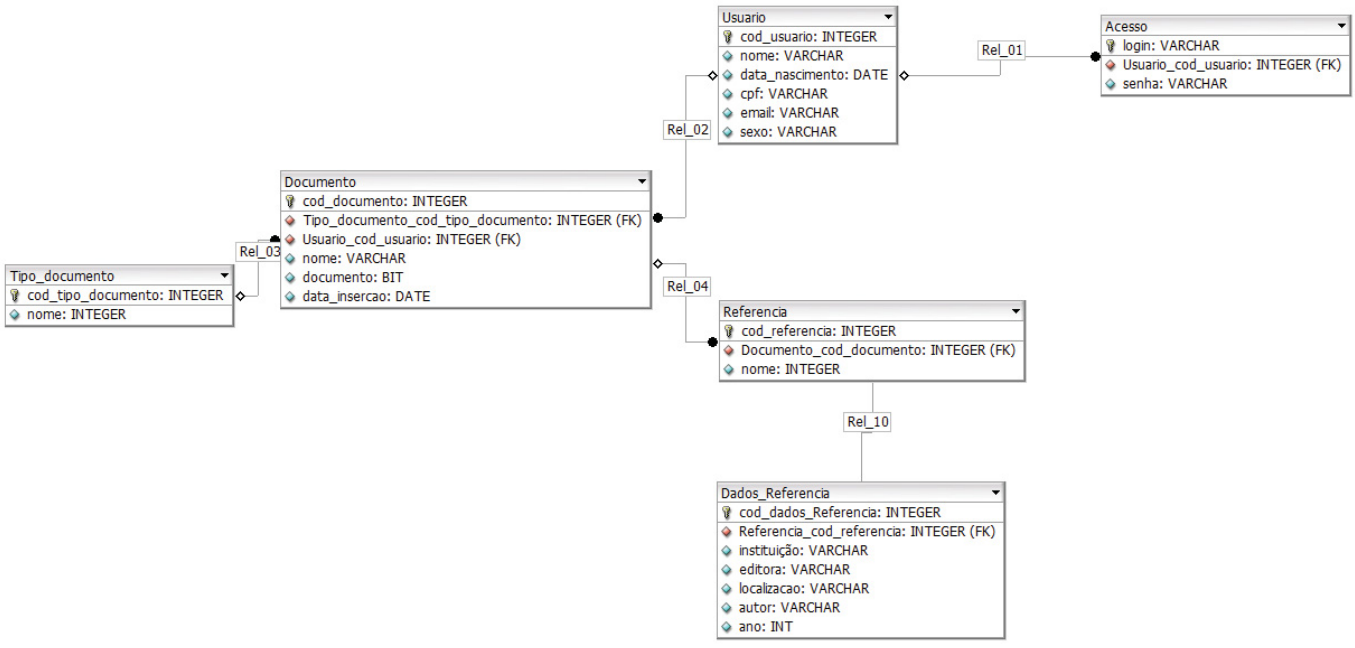
Na próxima seção, é apresentado o modelo de dados necessário para atender ao problema proposto.

1.2. MODELO DE DADOS

Nesta seção, é apresentado o modelo de banco de dados realizado para atender as demandas do sistema proposto, nele, são apresentadas as tabelas do banco que armazenarão informações coletadas e geradas pelo sistema.

A figura 33 apresenta este modelo:

Figura 33 – Modelo de Dados do Sistema



Fonte: Autor, 2015

No modelo de dados necessário para o sistema são apresentadas as seguintes tabelas:

- **Tipo de Documento:** Esta tabela armazena os tipos de documentos que serão inseridos no banco, podendo ser: teses, artigos, revistas, dentre outros. Após adicionar os documentos, o usuário tem a possibilidade de visualizá-los na lista de documentos adicionados;
- **Documento:** Esta tabela armazena os próprios documentos adicionados pelo usuário, separando as informações de nome do documento, a data de inserção e o próprio documento;
- **Usuário:** Nesta tabela, são armazenadas as informações de todos os usuários da plataforma, que são preenchidas na criação do perfil, ou na edição do mesmo;

- **Acesso:** Esta tabela armazena os dados de acesso de todos os usuários do sistema, sendo eles login e senha;

- **Referência:** Esta tabela armazena todas as referências coletadas do documento de forma íntegra.

- **Dados_Referencia:** Esta tabela armazena os dados extraídos das referências, armazenados de forma separada (autor, ano, etc).

Nesta seção, pôde-se visualizar o modelo do sistema proposto e o detalhamento das tabelas que compõem o modelo de dados.

5 DESENVOLVIMENTO DA SOLUÇÃO PROPOSTA

No presente capitulo, são apresentadas às tecnologias utilizadas para o desenvolvimento da aplicação proposta, descrevendo o método abordado e os resultados obtidos, além de uma apresentação do sistema desenvolvido.

5.1 FERRAMENTAS E TECNOLOGIAS

Nesta seção, são apresentadas as ferramentas e tecnologias escolhidas para o desenvolvimento do sistema proposto, apresentando uma descrição sobre o conceito de cada um das tecnologias e o motivo de sua escolha.

5.1.1 Apache OpenNLP

Apache OpenNLP é a sigla referente ao termo *Open Natural Language Processor*, em português, processamento de linguagem natural. Trata-se de um conjunto de ferramentas baseadas na aprendizagem de máquina para o processamento de texto de linguagem natural. (Apache OpenNLP, 2015).

O Apache OpenNLP possui algumas funções como quebra de textos, segmentação de frases, marcação, extração de entidade nomeada, detector de sentenças, dentre outras.

O objetivo do projeto OpenNLP será a criação de um kit de ferramentas madura para as tarefas anteriormente referidas. Um objetivo adicional é proporcionar um grande número de modelos pré-construído para uma variedade de línguas, assim como os recursos de texto anotado de que os modelos são derivadas. (Apache OpenNLP, 2015).

Apache OpenNLP (2015) afirma que, através dos componentes disponibilizados pela Apache OpenNLP, é possível construir um arsenal de processamento de linguagem natural completo.

Cada uma dessas ferramentas disponibilizadas pela Apache OpenNLP é acessível através de uma interface de programação de aplicativos (API). Além disso, uma interface de linha de comando (CLI) é fornecida para conveniência de experiências e formação.

5.1.1.1 Detector de Sentenças

O Detector de Sentenças OpenNLP, ferramenta fundamental para o desenvolvimento do sistema proposto como solução do problema, pode detectar que um caractere de pontuação marca o fim de uma frase. Neste sentido, uma sentença é definida como o maior espaço em branco aparado sequência de caracteres entre dois sinais de pontuação. O primeiro e último período fazer uma exceção a esta regra. O primeiro caractere não espaço em branco é considerado como o início de uma frase, e o último caractere não espaço em branco é assumido como sendo um fim frase. (Apache OpenNLP, 2015).

5.1.2 Plataforma JAVA

A plataforma JAVA é composta por um universo de ferramentas, na qual, no cenário de construção do sistema proposto, foi utilizada como a linguagem de programação para desenvolvimento do sistema.

> Java é uma linguagem de programação e plataforma computacional lançada pela primeira vez pela Sun Microsystems em 1995. Existem muitas aplicações e sites que não funcionarão, a menos que você tenha o Java instalado, e mais desses são criados todos os dias. O Java é rápido, seguro e confiável. De laptops a datacenters, consoles de games a supercomputadores científicos, telefones celulares à Internet, o Java está em todos os lugares. (JAVA, 2015).

Segundo Java (2015), esta tecnologia é a base para todos os tipos de aplicações em rede, sendo utilizados como padrão global para o desenvolvimento e distribuição de aplicações móveis, jogos, conteúdo baseado na Web e softwares corporativos.

A seguir são apresentados alguns números que comprovam isso, segundo Java (2015):

- 97% dos Desktops Corporativos executam o Java;
- 89% dos Desktops (ou Computadores) nos EUA Executam Java;
- 9 Milhões de Desenvolvedores de Java em Todo o Mundo;
- A Escolha Nº 1 para os Desenvolvedores;
- Plataforma de Desenvolvimento Nº 1;
- 3 Bilhões de Telefones Celulares Executam o Java;
- 100% dos Blu-ray Disc Players Vêm Equipados com o Java;
- 5 bilhões de Placas Java em uso;
- 125 milhões de aparelhos de TV executam o Java;
- 5 dos 5 Principais Fabricantes de Equipamento Original Utilizam o Java ME.

Java (2015) afirma que a ferramenta Java foi testada, refinada, estendida e comprovada por uma comunidade didática de desenvolvedores, arquitetos e entusiastas do Java, sendo projetado para permitir o desenvolvimento de aplicações portáteis de alto desempenho para uma ampla variedade de plataformas de computação

5.1.3 Servlet

Temple e outros (2004) definem Servlet como:

> Servlets são classes Java, desenvolvidas de acordo com uma estrutura bem definida, e que, quando instaladas junto a um Servidor que implemente um Servlet Container (um servidor que permita a execução de Servlets, muitas vezes chamado de Servidor de Aplicações Java), podem tratar requisições recebidas de clientes. (TEMPLE et al, 2004, p.11).

Temple e outros (2004) afirmam que, os Servlets recebem requisições do servidor que conseguem, através de sua tecnologia, capturar os parâmetros da requisição e realizar qualquer procedimento referente a uma classe Java, além de desenvolver uma pagina HTML. (TEMPLE, 2004).

5.1.4 Java Server Pages (JSP)

JSP é uma abreviação da sentença Java Server Pages e referem-se a um documento de texto que contém dois tipos de dados, os dados estáticos, que podem ser expressos em qualquer formato baseado em texto (HTML, SVG, WML e XML) e elementos vindos do JAVA que possibilitam a construção de um conteúdo dinâmico (Oracle, 2015).

Oracle (2015) afirma que:

> Tecnologia JavaServer Pages (JSP) fornece uma maneira simplificada, rápida para criar conteúdo web dinâmico. Tecnologia JSP permite o rápido desenvolvimento de aplicações baseadas na web que são de servidor e independente de plataforma.

As principais características da tecnologia JSP são, como se segue, segundo a Oracle (2015):

- A linguagem para o desenvolvimento de páginas JSP, que são documentos baseados em texto que descrevem como processar um pedido e construir uma resposta;
- Uma linguagem de expressão para acessar objetos do lado do servidor;
- Mecanismos para a definição de extensões para a linguagem JSP;

5.1.5 PostgreSQL

PostgreSQL é, segundo PostgreSQL (2015), um sistema poderoso de banco de dados *open source* (fonte aberta) objeto-relacional. Há mais de 15 anos em desenvolvimento

ativo o banco de dados PostgreSQL possui uma arquitetura comprovada, podendo ser executado em todos os sistemas operacionais, incluindo Linux e UNIX. (PostgreSQL, 2015).

O código fonte do PostgreSQL está disponível sob uma licença *open source* liberal. Esta licença dá a liberdade a qualquer pessoa para usar, modificar e distribuir PostgreSQL em qualquer forma, fonte aberto ou fechado. Como tal, o PostgreSQL não é apenas um sistema poderoso banco de dados capaz de executar o empreendimento, é uma plataforma de desenvolvimento evolutiva. (PostgreSQL, 2015).

5.2 HISTÓRICO DE DESENVOLVIMENTO

Para desenvolvimento do sistema como solução proposta para resolução do problema apresentado, a implementação esta separada em cinco etapas, sendo elas: análise inicial do sistema, implementação da extração das informações em texto, implementação do sistema, a geração de indicadores a partir das informações extraídas e, por ultimo, a fase de testes do sistema.

A primeira etapa a qual se chama de análise do sistema foi utilizada para idealizar como ficaria o sistema, o que precisaríamos fazer para desenvolvê-lo e a definição das demais etapas do sistema. A partir dela, foi pensado como seria desenvolvido o sistema, tanto visualmente (front-end), como funcionalmente (back-end), formalizando modelo de dados e a lógica de funcionamento do sistema. Ao finalizar esta etapa, iniciou-se o desenvolvimento a partir da extração de dados que é apresentada a seguir.

Para implementação da segunda etapa do sistema, foi necessário dedicar bastante tempo para estudo de ferramentas utilizadas para extração de informações em texto e testes para certificação do conteúdo extraído. Foi necessário procurar por ferramentas que pudessem auxiliar esta etapa, além de elaboração de termos de extração, utilizando expressões regulares. Nesta etapa, primeiramente, foi elaborada uma lógica para separar cada referência bibliográfica. A partir deste momento, iniciou-se o trabalho de quebra de informações das referencias. Uma ferramenta de grande importância nesta etapa foi o detector de sentenças da Apache OpenNLP. Com isso, facilitou-se a quebra de sentenças dos dados. Na maioria dos

casos, o detector de sentenças separava a referência em três sentenças, sendo elas: autores, título e subtítulo, e demais informações como editora, ano, revista, local, etc.

A partir dessa quebra, foi trabalhado a primeira e a última sentença para realizar a extração dos dados. Na primeira sentença, foi separado cada um dos autores da referência em questão e armazenados separadamente referenciando a qual documento estes autores pertencem. Na terceira sentença, foram aplicadas técnicas de expressões regulares para extrair o ano na qual aquela referência foi publicada e, assim, como nos autores, foi armazenada em um banco de dados relacional, referenciando a qual documento pertence este dado extraído.

Após se conseguir extrair essas informações, iniciou-se a terceira etapa, a de construção do sistema. Utilizando linguagem JAVA e as tecnologias de Servlet e JSP, desenvolveu-se o sistema contemplando todos os casos de usos, com exceção apenas da geração de indicadores, a qual foi realizada em uma etapa separada, apresentada a seguir.

Na quarta etapa do sistema, chamada de geração de indicadores, foi realizada a apresentação dos dados extraídos através da logica construída na segunda etapa. Para isso, realizamos consultas especificas no banco de dados, agrupando as informações de anos e autores extraídas, possibilitando a apresentação separada por documento, ou de todos os documentos adicionados pelo usuário no sistema. Em seguida, integraram-se as consultas ao sistema e criou-se a tela de apresentação das informações como uma barra de progresso para facilitar a visualização.

Por fim, a quinta etapa foi utilizada para realizar testes no sistema, tanto de funcionamento quanto de desempenho e credibilidade das informações extraídas. E, assim, finalizou-se a solução proposta.

5.3 SISTEMA DESENVOLVIDO

Nesta seção, é apresentado o funcionamento do sistema desenvolvido, mostrando o resultado das telas do sistema e suas funcionalidades.

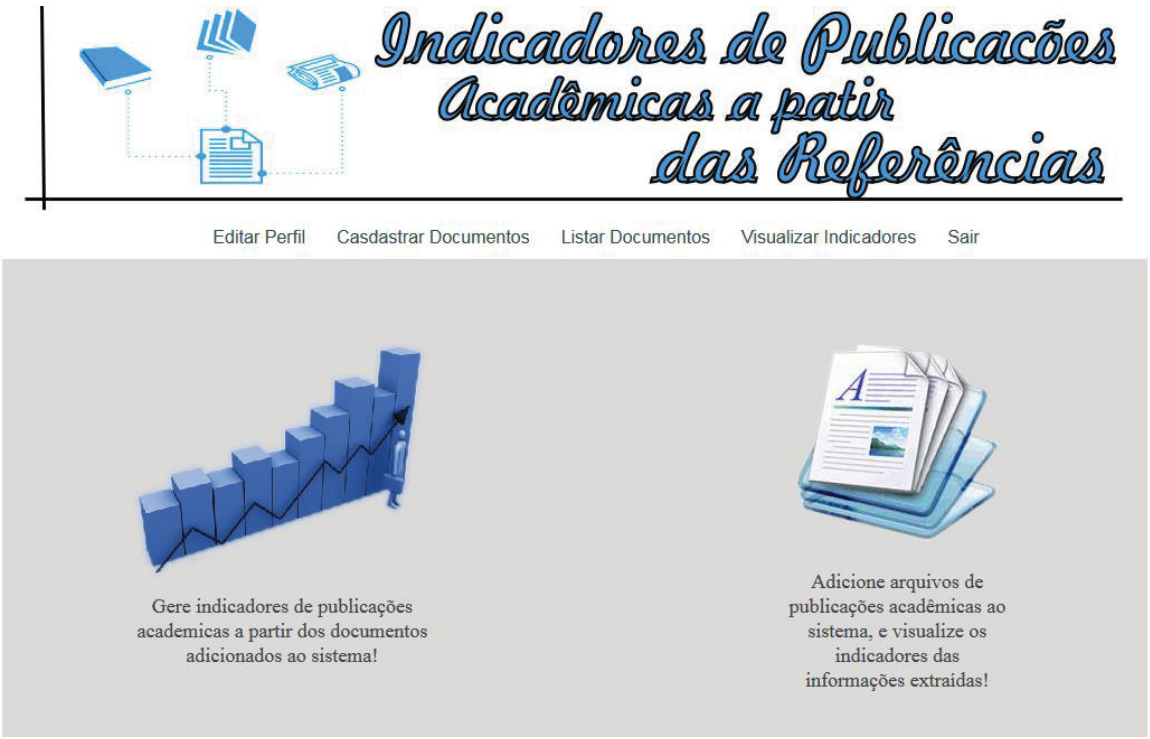
A figura 34, apresentada, ilustra a tela de login do sistema. Nesta etapa o usuário pode realizar seu login no sistema, caso já se tenha cadastrado, ou, então, em caso de primeiro acesso, realizar o cadastro no sistema.

Figura 34 – Tela de login do sistema

Fonte: Autor, 2015.

Como apresentado na figura 34, o usuário que já realizou o cadastro no sistema, após realizar o seu login, é direcionado para a tela inicial do sistema, conforme ilustrado na figura 35.

Figura 35 – Tela Inicial do Sistema



Fonte: Autor, 2015.

A tela inicial do sistema possui o objetivo de apresentar o sistema ao usuário com as imagens e textos contidos no seu conteúdo da pagina, além de direcionar o usuário para as funcionalidades disponibilizadas pelo sistema através de um menu superior localizado abaixo do cabeçalho do sistema.

Caso o usuário que acesse a tela de login do sistema, ainda, não tenha realizado o cadastro, é necessário que, então, o faça através da tela de cadastro de usuário, ilustrada pela figura 36.

Figura 36 – Tela de cadastro de usuário

Fonte: Autor, 2015.

Após realizar o cadastro do usuário, o sistema valida as informações e direciona o usuário para a tela principal do sistema, ilustrada pela figura 35.

Esta mesma tela de cadastro de usuário é utilizada para a edição do perfil, caso o usuário deseje alterar alguma informação, já cadastrada anteriormente.

Após acessar ao sistema, o usuário pode adicionar documentos ao sistema, esses documentos devem ser compostos por referências bibliográficas no formato da ABNT, norma na qual o sistema foi desenvolvido para extrair os dados.

O usuário pode adicionar os documentos através da tela de cadastro de documentos ilustrada pela figura 37.

Figura 37 – Tela de cadastro de documentos

Fonte: Autor, 2015.

Após realizar o cadastro de um documento, o sistema realiza a extração dos dados e armazena esses dados no banco de dados relacional PostgresSQL.

Ao cadastrar um documento, o usuário informa o nome deste documento, o tipo de arquivo (Artigo, Tese de Doutorado, TCC, etc.) e o próprio documento. Estas informações são salvas e podem ser visualizadas na tela de listagem dos documentos adicionados, ilustrada pela figura 38.

Figura 38 – Tela de lista de documentos adicionados

Fonte: Autor, 2015.

Além das informações inseridas na adição de documentos, a listagem apresenta também a data em que o documento foi inserido e uma coluna, chamada “Visualizar Indicadores”, na qual é disponibilizado um link para a visualização dos indicadores referente ao documento que está sendo apresentado na listagem. Os indicadores disponibilizados são por autores e ano de publicação. É possível visualizar a tela de indicadores através das figuras 40 e 41, respectivamente.

Os indicadores podem ser visualizados também pela opção de menu chamada “Visualizar Indicadores” contida no menu do sistema. Ao clicar nesta opção de menu, o sistema direciona o usuário para a tela na qual é possível escolher o tipo de relatório a ser gerado e o documento no qual se deseja visualizar os dados. É possível gerar um relatório, contemplando todos os documentos adicionados pelo usuário, também, conforme apresentado na figura 39.

Figura 39 – Tela de filtro para geração de indicadores

Fonte: Autor, 2015.

Após selecionar a opção de relatório desejada e o documento para visualização dos dados, o sistema apresenta as informações, conforme ilustrado pelas figuras 40 e 41. Nela, é apresentado o dado em questão (autor ou ano), a quantidade agrupada por documento e uma barra de progressão para facilitar a visualização.

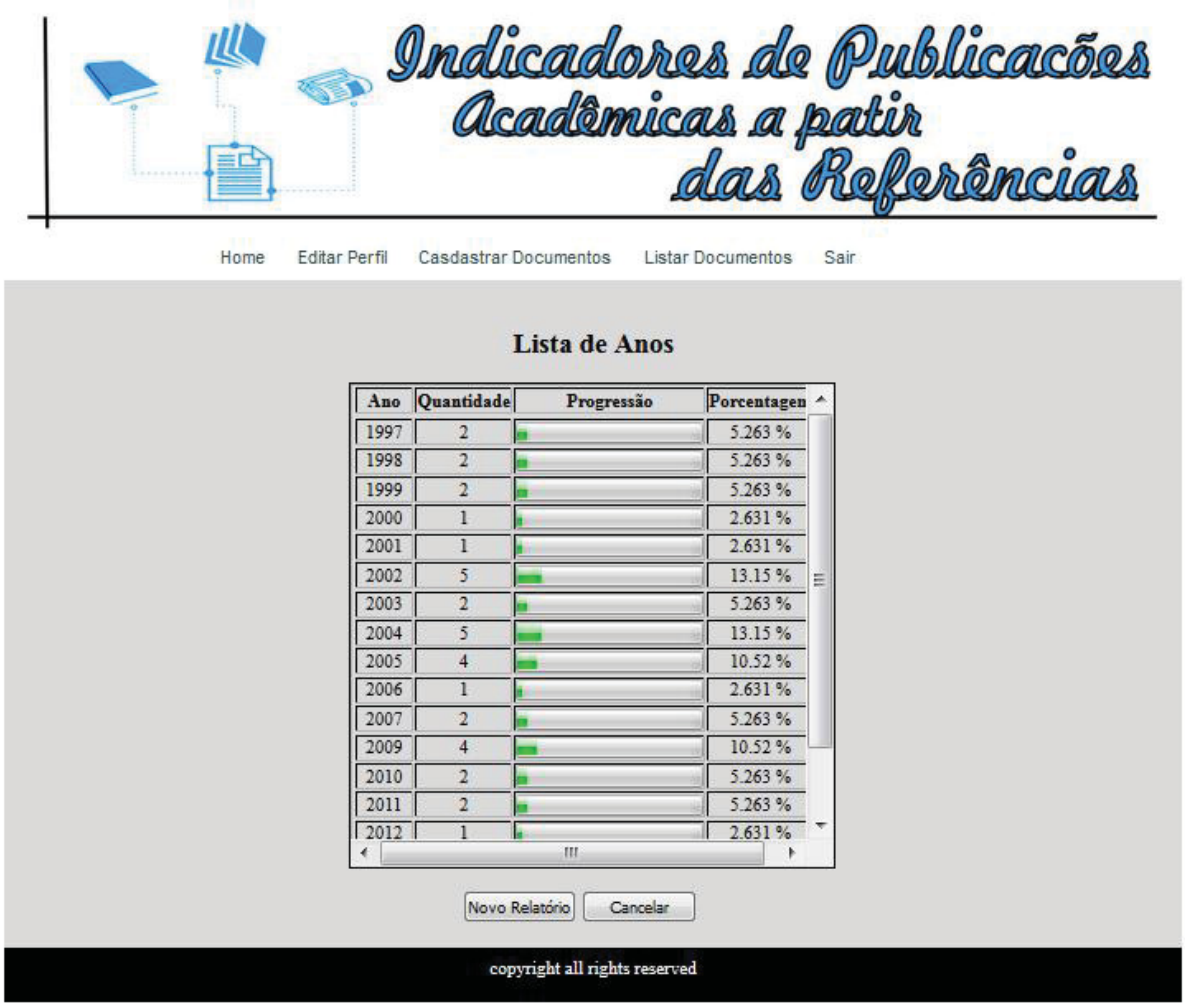
Figura 40 – Tela de visualização de indicadores de autor

Fonte: Autor, 2015.

Como é possível ver na figura 40, a tabela apresenta todos os autores extraídos e a quantidade encontrada. Por exemplo: O autor DANIEL SCHWABE foi utilizado uma vez no documento pesquisado. Sendo que este significa 1,538 % em relação ao todo. Sendo assim, 1,538% de todos os autores utilizados são de DANIEL SCHWABE.

A figura 41 ilustra os mesmos dados, entretanto, realizando a pesquisa por ano de publicação.

Figura 41 – Tela de visualização de indicadores de anos



Fonte: Autor, 2015.

Como é possível visualizar, através da figura 41, assim como na figura 40, o sistema apresenta uma tabela, mostrando o ano de publicação, a quantidade de referências utilizadas referente aquele ano e uma barra de progressão, e este valor apresentado em porcentagem em relação ao todo. Por exemplo: no documento pesquisado, o ano no qual se encontrou mais referências bibliográficas foi o ano de 2002, contendo 5 registros cada, isso resume em 13,15% do total da pesquisa realizada, ou seja, 13,15% das referencias utilizadas são do ano de 2002.

Nesta seção, foi possível ver o resultado do sistema desenvolvido, contendo todas as suas funcionalidades, e o resultado dos dados já adicionados ao sistema.

5.4 AVALIAÇÃO DO SISTEMA

O sistema proposto para solução do problema explícito no inicio do documento foi avaliado através do percentual de acerto que o sistema apresentou com a extração da informação e apresentação dos dados. Para isso, foi comparado o valor apresentado pelo o sistema com os documentos originais, chegando assim, a um percentual de acertos e erros.

O desenvolvimento do sistema trabalhou sobre duas informações de referências bibliográficas, sendo elas: autores, e ano de publicação. Dessa forma, a avaliação é feita separada para cada um dos dados extraídos.

5.4.1 Desempenho da extração de autores

Em relação aos autores, para avaliar o desempenho do sistema de acordo com o que foi desenvolvido, foram adicionadas 535 referências de 11 arquivos científicos diferentes, variando entre artigos científicos e trabalhos de conclusão de curso de monografias já avaliadas pela própria instituição (UNISUL).

Dessas 535 referencias foram extraídos 735 autores, sendo que, o total de autores contidos nestes documentos era de 782, ou seja, o sistema não conseguiu recuperar 47 de todos os autores contidos nos documentos.

Destes 735 autores extraídos pelo sistema, 24 deles possuem alguma inconsistência, ou seja, a informação não foi recuperada de forma totalmente correta, como se pode ver através da figura 42.

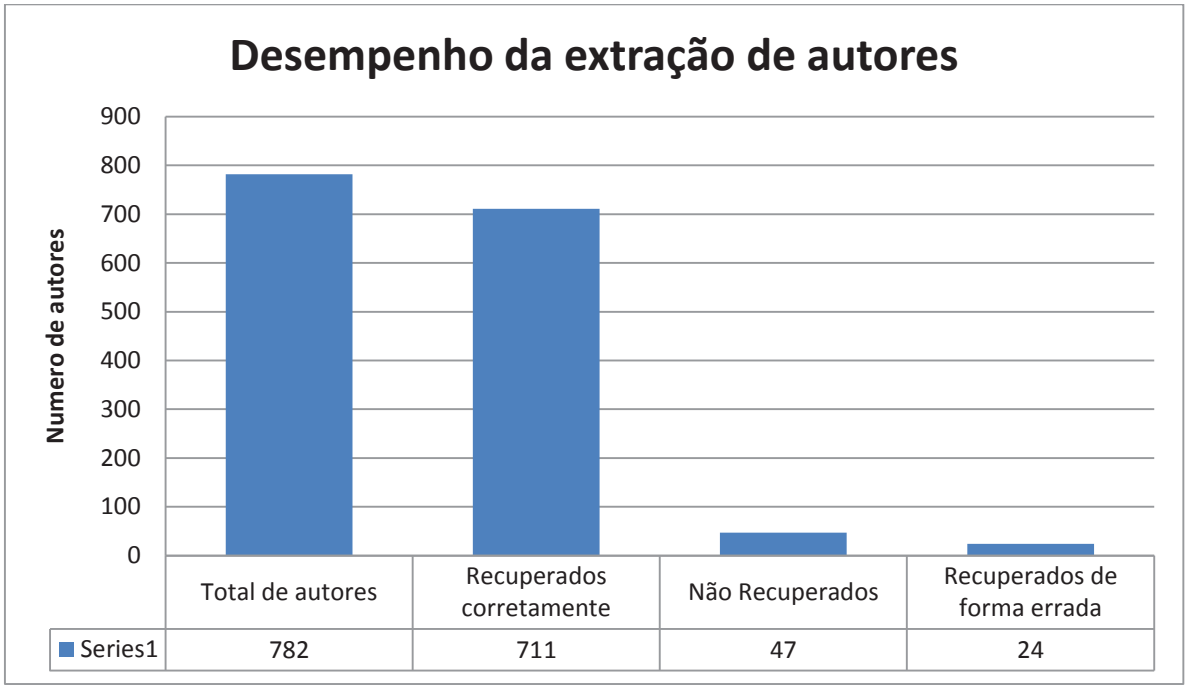
Figura 42 – Dados recuperados de forma errônea.

Fonte: Autor, 2015.

Sendo assim, o sistema recuperou de forma totalmente correta 711 autores de um total de 782, resultando em um percentual de acerto de 90,92 %.

A figura 43, apresentada a seguir, possibilita uma melhor visualização dos resultados.

Figura 43 – Desempenho da extração de autores.



Fonte: Autor, 2015.

Após uma análise minuciosa dos 9,18% dos autores não recuperados ou recuperados de forma errada, foi constatado que, na sua maioria, os erros ocorreram por mal formulação das referências bibliográficas, e não por falha do sistema.

Como podemos ver na expressão a seguir, no exemplo apresentado na figura 42, a referência bibliográfica estava formulada da seguinte maneira:

BLATTMANN, U., SILVA, F. C. C. (2007). **Colaboração e interação na web 2.0 e biblioteca 2.0**. Revista ACB: Biblioteconomia em Santa Catarina, vol. 12 nº 2, 191-215.

Assim após o a extração dos autores, o sistema recuperou a seguinte expressão: “BLATTMANN, U., SILVA, F. C. C. (2007)”.

Caso a referencia estivesse formulada de forma correta, como no exemplo apresentado a seguir:

BLATTMANN, U.; SILVA, F. C. C. **Colaboração e interação na web 2.0 e biblioteca 2.0**. Revista ACB: Biblioteconomia em Santa Catarina, vol. 12 nº 2, 2007, p.191-215.

O sistema teria recuperado os dois autores contidos na referência, sendo eles: U. BLATTMANN e F. C. C. SILVA.

Como foi explicado no decorrer desse documento, o sistema foi desenvolvido para trabalhar sobre uma norma de elaboração de referências, sendo ela a ABNT. A regra define um padrão para formulação das referencias com os elementos em seus devidos lugares e separadores, quando esse padrão não é respeitado, o sistema acaba deixando de recuperar informações importantes como mostrado no exemplo anterior.

5.4.2 Desempenho da extração de anos

Referente à extração do ano de publicação das referencias, pode-se afirmar que, houve uma maior facilidade na recuperação dos dados em relação aos autores por se trabalhar com expressões regulares.

Das 535 referências adicionadas ao sistema, foram recuperadas de forma totalmente correta 523 registros. Sendo que, duas referências estavam sem o ano de publicação ou ano de acesso e as demais não foram recuperadas.

A maior dificuldade para extração do ano de publicação se da quando a referência apresenta mais que um ano em sua formulação, geralmente, contendo datas de acesso a sites da WEB, como na referência apresentada a seguir, onde se encontram os anos de 2005 e 2012:

O'REILLY, Tim. What is web 2.0. **Design patterns and business models for the next generation of software**. 2005. Disponível em: http://facweb.cti.depaul.edu/jnowotarski/se425/What%20Is%20Web%202%20point%200.pdf . Acessado em: 20 set. 2012.

Para estes casos, foi considerado como ano de publicação o registro mais antigo, neste caso, o ano de 2005.

Muitas das referencias não possuíam ano de publicação, como no exemplo a seguir:

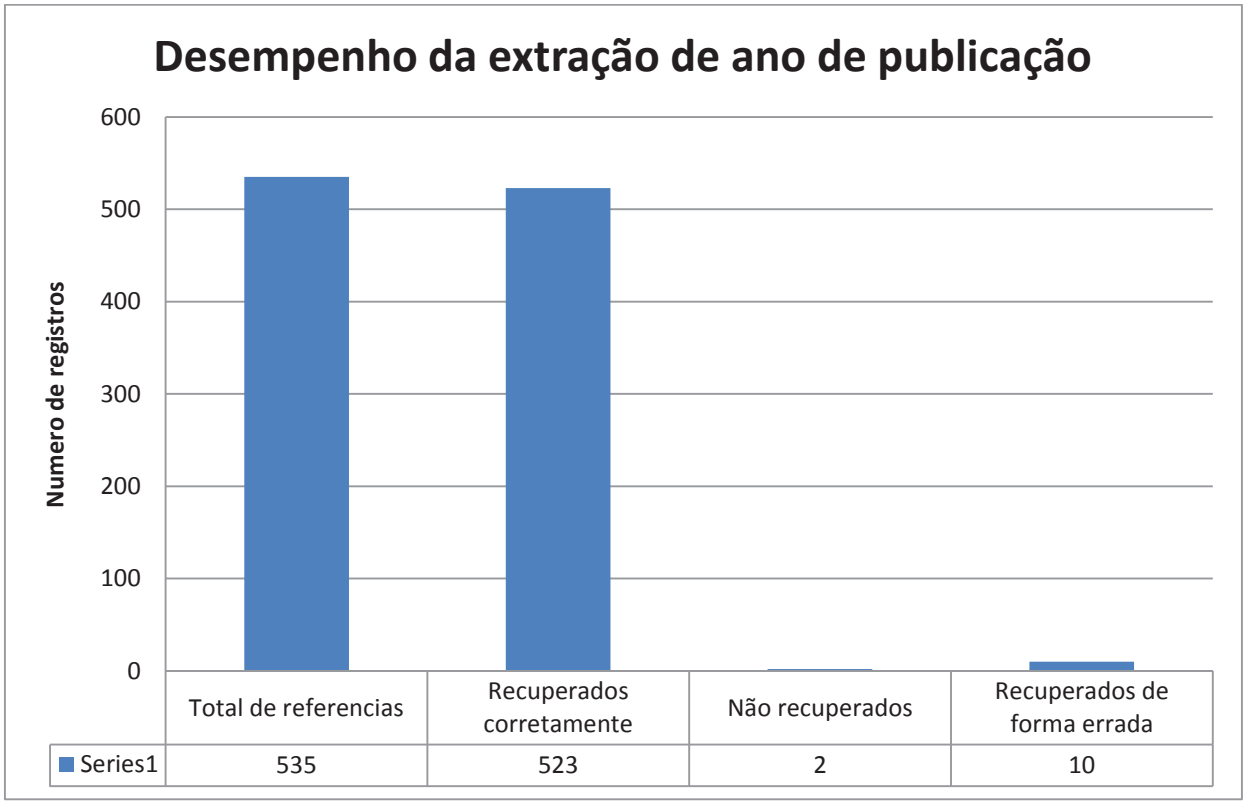
DENATRAN. **Frota de veículos no Brasil**. Disponível em: http://www.denatran.gov.br/frota.htm. Acesso em: 29 de ago. 2012.

Nestes casos, foi considerado como ano de publicação, o ano de acesso ao site, ou seja, como apresentado no exemplo anterior o ano considerado foi 2012.

Desta forma, o percentual de acertos apresentado pelo sistema em relação ao ano de publicação foi de 97,75%.

A figura 44 apresentada representa melhor o resultado obtido:

Figura 44 - Desempenho da extração de ano de publicação



Fonte: Autor, 2015.

6 CONCLUSÕES E TRABALHOS FUTUROS

No presente capitulo, são apresentados os resultados obtidos com o desenvolvimento da aplicação WEB para atingir o objetivo principal especificado. São apresentados também, os trabalhos futuros referentes ao tema e os problemas encontrados no decorrer do desenvolvimento.

6.1 CONCLUSÃO

No decorrer deste trabalho foi apresentado conceitos sobre extração de informações em texto e indicadores de publicações acadêmicas com o objetivo principal de representar estes conceitos em forma de um sistema que realiza a leitura de referências bibliográficas em forma de texto e apresentar informações agrupadas sobre as mesmas.

Para desenvolvimento da pesquisa, utilizaram-se duas principais tecnologias que, foram essenciais para que a proposta conseguisse alcançar o objetivo descrito anteriormente. O detector de sentenças, que foi utilizado para realizar a quebra das informações contidas nas referências bibliográficas, e o uso das expressões regulares atendendo a etapa de reconhecimento de padrão que possibilitou encontrar os objetos necessários para extração.

Para avaliar a aplicação desenvolvida foi realizada uma comparação manual das informações contidas nos documentos adicionados ao sistema, com a extração dos dados apresentados pelo mesmo.

A principal dificuldade encontrada foi a má formulação de referências bibliográficas nos documentos utilizados como fonte de dados para o sistema. Além do fato de se trabalhar com uma área desde então desconhecida pelos autores deste trabalho.

Em relação aos problemas apresentados no início do trabalho, através do sistema desenvolvido como proposta de solução, e, os resultados obtidos com o mesmo, pode-se concluir que a proposta é valida e atende completamente o problema descrito.

Com a análise dos resultados obtidos com o desenvolvimento da aplicação, pode-se afirmar que, o sistema atende, na sua grande maioria, os objetivos apresentados. Entretanto, aplicado sobre duas principais informações contidas nas referencias bibliográficas, sendo elas, os autores e o ano de publicação das referências, e não sobre todos os elementos nela contidos.

6.2 TRABALHOS FUTUROS

Em relação aos trabalhos futuros, os autores têm como principal objetivo, o aperfeiçoamento na extração das informações contidas nas referências bibliográficas, ou seja, recuperar um maior número de informações possíveis, como: local de publicação, periódicos, editoras, etc.

Tem-se como objetivo também, estender os padrões de leitura de referências, aceitando padrões internacionais como Chicago e Harvard, por exemplo. Considerando que atualmente o sistema realiza leitura de referencias escritas nos padrões especificados pela ABNT.

Além disso, os autores tem a intenção de integrar uma tecnologia de gráficos ao sistema que possa representar melhor os resultados obtidos na geração de indicadores.

Uma proposta para trabalhos futuros, também, é, aumentar o número de extensões de arquivos aceitas pelo sistema, possibilitando assim, documentos em diversos outros formatos para a realização da leitura e extração dos dados.

REFERÊNCIAS

ALMEIDA, Mauricio Barcillos. **Uma introdução ao XML:** Sua utilização na Internet e alguns conceitos complementares. Dissertação (Mestrado) - Curso de Ciência da Informação, Universidade Federal de Minas Gerais, Brasília, 2002, p.4.

ALVARENGA, Lídia. **Bibliometria e arqueologia do saber:** de Michel Foucault-traços de identidade teórico-metodológica. TCC (Graduação) - Curso de Ciência da Informação, Universidade Federal de Minas Gerais, Brasília, 1998.

ÁLVAREZ, Alberto Cáceres. **Extração de informação de artigos científicos**: uma abordagem baseada em indução de regras de etiquetagem. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2007.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6023**: Informação e documentação: trabalhos acadêmicos: apresentação. Rio de Janeiro, 2002.

Apache OpenNLP. Página oficial do Apache OpenNLP. 2015, Disponível em: <https://opennlp.apache.org>, Acesso em: Outubro, 2015.

BATISTA, Carlos Eduardo C. F.; SCHWABE, Daniel. **LinkedTube: Informações Semânticas em Objetos de Mídia da Internet**. Simpósio Brasileiro de Sistemas Multimídia e Web, Fortaleza, 2009, p.15.

BORDIN, Andréa; CECI, Flavio; GONÇALVES, Alexandre; GAUTHIER, Fernando; PACHECO, Roberto. **Análise da Produção Científica do Simpósio Internacional sobre Interdisciplinaridade no Ensino**. SIMPÓSIO INTERNACIONAL SOBRE INTERDISCIPLINARIDADE NO ENSINO NA PESQUISA E NA EXTENSÃO–REGIÃO SUL, Florianópolis, 2013, p.3.

BOOCH, Grady; RUMBAUGH, James; JACOBSON, Ivar. **UML: guia do usuário**. Editora Elsevier Brasil, São Paulo, 2006, p.7.

CABRAL, Luís. **Sistema Uniformizado de Pesquisa de Referências Bibliográficas.** Dissertação (Mestrado) - Curso de Mestrado de Engenharia Informática, Faculdade de Engenharia da Universidade do Porto, Porto, 2005.

CARVALHO, Ana E.; TAVARES, H. C. A. B.; CASTRO, Jaelson. **Uma Estratégia para Implantação de uma Gerência de Requisitos visando a Melhoria dos Processos de Software**. Buenos Aires, Argentina, 2001.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR. **Classificação da produção intelectual**. 2014. Disponível em: <http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/classificacao-da-producao-intelectual>. Acesso em: Março, 2015.

DA SILVA, Vanessa Gomes; DA SILVA, Maristela Gomes; AGOPYAN, Vahan. **Avaliação de edifícios: definição de indicadores de sustentabilidade**. III ENECS – Encontro nacional sobre edificações e comunidades sustentáveis, 2003.

DE PÁDUA, Silvia Inês Dallavalle. **Investigação do Processo de Desenvolvimento de Software a partir da Modelagem Organizacional, enfatizando regras do negócio**. Tese de Doutorado, Universidade de São Paulo, São Carlos, 2001.

FERREIRA, Helder Filipe PatrÍcio Cabral. **Leitura automática de expressões matemáticas – audio math.** Dissertação (Mestrado) - Curso de Departamento de Engenharia e Eletrotécnica e de Computadores, Faculdade de Engenharia da Universidade do Porto, Porto, 2005.

FOWLER, Martin. **UML Essencial:** Um breve guia para linguagem padrão. Editora Bookman, Porto Alegre, 2005.

FURTADO, Maria Inês Vasconcellos**. Inteligência Competitiva para o Ensino Superior Privado: Uma abordagem através da mineração de textos**. Dissertação (Mestrado) Universidade Federal do Rio de Janeiro, 2004.

GOYVAERTS, Jan; LEVITHAN, Steven. **Expressões Regulares Cookbook.** Novatec Editora, São Paulo, 2011. 27 p.

INSTITUTO DE RELAÇÕES INTERNACIONAIS – USP. **Citação e referência bibliográfica em formatos ABNT e Chicago**. 16º edição, São Paulo, 2012.

JARGAS, Aurelio Marinho. **Expressões Regulares: uma abordagem divertida.** 3 ed., Novatec Editora, São Paulo, 2009. 120 p. 4.

JAVA. **Página oficial da plataforma Java**. 2015, Disponível em: <http://www.java.com/pt_BR/about/>, Acesso em: Outubro, 2015.

KAMIENSKI, C.A. **Introdução ao paradigma de Orientação a Objetos.** Centro Federal de Educação Tecnológica da Paraíba Diretoria de Ensino. João Pessoa, 1996.

KAUARK, Fabiana da Silva; MANHÃES, Fernanda Castro; MEDEIROS, Carlos Henrique. **METODOLOGIA DA PESQUISA:** UM GUIA PRÁTICO. Litterarum Editora, Itabuna, 2010.

LEONEL, Vilson; MOTTA, Alexandre de Medeiros. **Ciência e Pesquisa.** Universidade do Sul de Santa Catarina (UNISUL), 2 ed., Palhoça, 2007. 221 p.

LOH, Stanley; DE OLIVEIRA, José Palazzo M. **Descoberta de conhecimento em textos.** TCC (Graduação) - Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 1999.

LOH, Stanley; GARIN, Ramiro Saldaña; **Web Intelligence–Inteligência Artificial para Descoberta de Conhecimento na Web**. Oficina de Inteligência Artificial, Editora da UCPEL , Pelotas RS, v.5, 2001, p. 11-34.

LOH, Stanley; WIVES, Leandro Krug; DE OLIVEIRA, José Palazzo M. **Descoberta proativa de conhecimento em textos: aplicações em inteligência competitiva.** Simpósio Internacional da Gestão do Conhecimento, 2000.

LOPES, Maria C.S. **Mineração de dados textuais utilizando técnicas de clustering para o idioma português**, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

MAIA, José Anízio. **Construindo softwares com qualidade e rapidez usando ICONIX**. 2005. Disponível em: <http://www.guj.com.br/content/articles/patterns/iconix_guj.pdf>

MIRANDA, Roberto Campos da Rocha. **O uso da informação na formulação de ações estratégicas pelas empresas.** Monografia (Especialização) - Curso de Ciência da Informação, Brasília, 1999.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. **Conceitos sobre aprendizado de máquina:** Sistemas Inteligentes-Fundamentos e Aplicações. v. 1, p. 1, 2003.

MUGNAINI, Rogério; JANNUZZI, Paulo; QUONIAM, Luc. **Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base Pascal**; Ciência da Informação, 2004.

NARDI, Julio Cesar; FALBO, Ricardo de Almeida. **Uma Ontologia de Requisitos de Software**. 2006.

ORACLE. **Página Oficial da tecnologia de trabalho JSP**. 2015, Disponível em: <http://www.oracle.com/technetwork/java/javaee/jsp/index.html>, Acesso em: Outubro, 2015.

PRODANOV, Cleber Cristiano; DE FREITAS, Ernani Cesar. **Metodologia do Trabalho Científico**: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico. Editora Feevale, 2013.

PAPA FILHO, Sudário; VANALLE, Rosângela M. **O uso da informação como recurso estratégico de tomada de decisão**. Encontro Nacional de Engenharia de Produção, v. 22, Curitiba, 2002.

PostgreSQL. **Página oficial da plataforma PostgreSQL**. 2015, Disponível em: <http://www.postgresql.org/about/>, Acesso em: Outubro, 2015.

RAUPP, Fabiano Maury; BEUREN, Ilse Maria**. Metodologia da pesquisa aplicável às ciências sociais: Como elaborar trabalhos monográficos em contabilidade: teoria e prática**. 2003.

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

RINO, Lucia Helena Machado; SENO, Eloize Rossi Marques; **A importância do tratamento co-referencial para a sumarização automática de texto.** TCC (Graduação) - Curso de Departamento de Computação – Ufscar, Universidade Federal de São Carlos, São Paulo, 2006.

SCARINCI, Rui Gureghian. **Sistema de Extração Semântica de informações**. Dissertação (Mestrado) - Curso de Instituto de Informática Curso de Pós-graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1997.

SILVA FILHO, Luiz Alberto da. **Mineração de regras de associação utilizando KDD e KDT:** UMA APLICAÇÃO EM SEGURANÇA PÚBLICA. Dissertação (Mestrado) - Curso de Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2009, p 16.

SILVA, Cassiana Fagundes; VIEIRA, Renata; OSÓRIO, Fernando Santos. **Uso de Informações Linguísticas em Categorização de Textos utilizando Redes Neurais Artificiais.** VIII Simpósio Brasileiro de Redes Neurais, 2005.

SILVA, Adriano José da; FILHO, Jorge Ribeiro Toledo; PINTO, Juliana. **Análise bibliométrica dos artigos sobre controladoria publicados em periódicos dos programas de pós-graduação em ciências contábeis recomendados pela CAPES.** Revista ABCustos-Associação Brasileira de Custos, São Paulo, 2009.

SILVA, George et al. Utilizando ICONIX no desenvolvimento de aplicações DELPHI. II Congresso de Pesquisa e Inovação da Rede Norte Nordeste de Educação Tecnológica, João Pessoa-PB, 2007.

STREHL, Letícia. **O fator de impacto do ISI e a avaliação da produção científi ca: aspectos conceituais e metodológicos**. Ciência da Informação, v. 34, n. 1, Brasília, 2005, p. 19-27.

TELINE, Maria Fernanda. **Avaliação de Métodos de Extração Automática de Terminologia para textos em Português.** Ciências da Computação e Matemática Computacional, Instituto de Ciências Matemáticas e de Computação – ICMC-USP, Pós Graduação, São Paulo, 2004, p. 39-40.

TEMPLE, André et al**. Programação Web com Jsp, Servlets e J2EE**. v. 183, 2004. Disponível em: <http://www.icmc.usp.br/~mello/livro-j2ee.pdf>, Acesso em: Outubro, 2015.

THOMSON REUTERS. Journal Citation Report. **2010 JCR Science Edition**. New York, 2015. Disponível em: <http://wokinfo.com/products_tools/analytical/jcr/>, Acesso em: Março, 2015.

VANTI, Nadia Aurora Peres. **Da bibliometria à webometria:** uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento**.** Ciência da Informação, 2002.

VIEIRA, Valter Afonso. **As tipologias, variações e características da pesquisa de marketing**. Revista da FAE, Curitiba, 2002.

WIVES, Leandro Krug. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos.** Tese (Doutorado) - Curso de Instituto de Informática Curso de Pós-graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

WIVES, Leandro Krug; LOH, Stanley. **Recuperação de Informações usando a expansão Semântica e a Lógica Difusa.** CONGRESSO INTERNACIONAL EM INGENIERIA INFORMATICA, Universidad de Buenos Aires, Buenos Aires ,1998.

WIVES, Leandro Krug. **Um estudo sobre técnicas de recuperação de informações com ênfase em informações textuais.** Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 1997.