UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DEPARTAMENTO DE INFORMÁTICA
MESTRADO EM INFORMÁTICA

KARLA SAMANTHA BEZERRA VALE

AMBIENTE INTELIGENTE E COLABORATIVO PARA APOIO À PRODUÇÃO ACADÊMICA – ESCLARECIMENTO DE DÚVIDAS

Vitória
Fevereiro, 2015

KARLA SAMANTHA BEZERRA VALE

AMBIENTE INTELIGENTE E COLABORATIVO PARA APOIO À PRODUÇÃO ACADÊMICA – ESCLARECIMENTO DE DÚVIDAS

Dissertação submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Informática.

Vitória
Fevereiro, 2015

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial Tecnológica,
Universidade Federal do Espírito Santo, ES, Brasil)

Vale, Karla Samantha Bezerra, 1986-

V149s Ambiente inteligente e colaborativo para apoio à produção acadêmica : esclarecimento de dúvidas / Karla Samantha Bezerra Vale. – 2015.
95 f. : il.

Orientador: Crediné Silva de Menezes.

Dissertação (Mestrado em Informática) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Inteligência artificial – Aplicações educacionais. 2. Recuperação da informação. 3. Sistemas de consultas e respostas. I. Menezes, Crediné Silva de. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 004

KARLA SAMANTHA BEZERRA VALE

AMBIENTE INTELIGENTE E COLABORATIVO PARA APOIO À PRODUÇÃO ACADÊMICA – ESCLARECIMENTO DE DÚVIDAS

Dissertação submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Informática.

BANCA EXAMINADORA

_____
Prof. Dr. Crediné Silva de Menezes
Universidade Federal do Espírito Santo (UFES)

(Orientador)

_____
Prof. Dr. Orivaldo de Lira Tavares
Universidade Federal do Espírito Santo (UFES)

_____
Profª. Drª. Tânia Barbosa Salles Gava
Universidade Federal do Espírito Santo (UFES)

_____
Prof. Dr. Alberto Nogueira de Castro Junior
Universidade Federal do Amazonas (UFAM)

Vitória
Fevereiro, 2015

DEDICATÓRIA

À Tia Raimunda
(in memoriam)

AGRADECIMENTOS

Antes de tudo, agradeço a Deus, por ter sido minha fonte de calmaria e conforto nos momentos difíceis e, principalmente, pelo dom da vida.

Aos meus pais, Francisco das Chagas Vale e Maria Zélia Bezerra Vale, por terem apoiado as minhas decisões, pelo amor incondicional, pelos sacrifícios, conselhos e palavras de encorajamento. Em especial à minha mãe, por ser sempre a maior incentivadora da busca incessável de conhecimento e melhoria, meu maior exemplo.

À minha irmã, Jéssica Caroline Bezerra Vale, por ser a melhor irmã do mundo. Pela paciência, companheirismo, cumplicidade, incentivo, amor, ouvidos nos momentos de nervosismo e, até, raiva. Por ser a melhor pessoa que eu conheço e por não ter me deixado desistir.

À Luna que, mesmo sem entender, sempre proporcionou carinho e companhia sendo nos latidos matutinos ou nas lambidas pós-almoços.

Aos meus amigos que, direta ou indiretamente, contribuíram para a conclusão deste trabalho, nos momentos de descontração e de concentração. Em especial à Aldo Matos, Iara Silveira, Robson Azevedo, Sheldon Henrique, Stefano Henrique, Stanley Rodrigues, Maria Silva, Lucrécia Marques, Heloiza Merlin, Thábata Brito, Pedro Henrique Pantoja e Leonardo Medeiros, por terem sido pacientes e compreensivos com os meus rompantes e ausência.

Ao meu orientador, Prof. Dr. Crediné Silva de Menezes, pela infinita paciência nos momentos de dúvidas, pelas broncas necessárias e principalmente pelo direcionamento e aprendizado.

Aos demais professores do Programa de Mestrado em Informática, por serem referência de paixão da profissão e conhecimento.

À Universidade Federal do Espírito Santo, pela oportunidade de participar do programa de mestrado e pela estrutura fornecida.

À CAPES, por ter, gentilmente financiado parcialmente a produção deste trabalho.

One child,
one teacher,
one book and
one pen can change the world.

Malala Yousafzai

RESUMO

A elaboração de trabalhos acadêmicos, como sabemos, exige grande esforço dos autores, em grande parte de cunho operacional, ocupando horas de trabalho que poderiam ser dedicadas às atividades de análise e criação. Entre tais atividades podemos citar o levantamento de informações sobre o tema a ser pesquisado, o gerenciamento de artigos lidos ou a serem lidos, ou a busca por publicações relacionadas ao tema. Sabemos ainda que muitos destes esforços podem ser atenuados através de inciativas de cooperação e uso de ferramentas computacionais. Existem ferramentas conceituadas que oferecem suporte para algumas dessas etapas, no entanto o acadêmico precisa combinar duas ou mais dessas ferramentas para atender as suas necessidades. Procurando contribuir para atenuar esse problema concebemos no LIEd um Ambiente Inteligente e Colaborativo, que combina técnicas de Inteligência Artificial no intuito de proporcionar um ambiente colaborativo capaz de apoiar computacionalmente, de forma integrada, algumas das etapas essenciais da produção acadêmica. Este trabalho apresenta um subsistema deste ambiente, o Agente de Dúvidas, que apresenta informações de forma direta e automática a partir de perguntas em linguagem natural, considerando o contexto de um projeto e a base de documentos indicada pelo usuário durante as suas interações.

Palavras-chave: Inteligência Artificial, Recuperação de Informação, Sistemas de Esclarecimento de Dúvidas

ABSTRACT

The development of academic work, as we know, requires great effort of the authors, largely operational nature, occupying hours of work that could be dedicated to the analysis and creation activities. Among these activities we can mention the collection of information on the subject to be searched, the management items read or to be read, or the search for publications on the theme. Yet we know that many of these efforts can be mitigated through initiatives of cooperation and use of computational tools. There are reputable tools that support some of these steps, however the academic needs to combine two or more of these tools to attend his needs. Looking help to alleviate this problem we designed, at LIEd, an Intelligent and Collaborative Environment, which combines artificial intelligence techniques in order to provide a collaborative environment able to support computationally, in an integrated way, some of the essential steps of the academic production. This paper presents a subsystem of the environment, Document Retrieval, which provides information directly and automatically from questions in natural language, considering the context of a project and the documentary evidence indicated by the user during their interactions.

Keywords: Artificial Intelligence, Information Retrieval, Questions Answering Systems

LISTA DE SIGLAS

AIML	Artificial Intelligence Markup Language
API	Application Programming Interface
CLEF	Conference and Labs of the Evaluation Forum
EI	Extração de Informação
FAQ	Frequently Asked Questions
IA	Inteligência Artificial
IDC	International Data Corp
MVC	Model-View-Controller
NLTK	Natural Language ToolKit
PLN	Processamento de Linguagem Natural
QAS	Question Answering Systems
RI	Recuperação de Informação
RDF	Resource Description Framework
TREC	Text REtrieval Conference
URL	Uniform Resource Locator
XML	eXtensible Markup Language

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em tempos do aumento na disponibilização de conteúdo, popularização das redes sociais e o crescente compartilhamento de informações, o grande volume e a diversidade de informações disponíveis na Internet oferecem novas oportunidades de acesso e de exploração desse conteúdo.

Segundo REDDY; SCIENCE; STATE (2010), a tecnologia da comunicação desempenha um papel vital no desenvolvimento da sociedade. Vastas quantidades de dados são transmitidos em segundos, e acesso à Internet oferece grandes quantidades de informação, dados e materiais interpretados. Como uma ferramenta poderosa e dinâmica para a comunicação, é a maior fonte de informação a nível global. Considerando esse espaço virtual como um grande repositório de informação, faz-se necessário a criação de mecanismos de busca, capazes de lidar com essa grande quantidade de dados de forma relevante. Mesmo com os atuais portais de busca existem aspectos dessas ferramentas que podem ser explorados sob um determinado contexto.

Em LARSEN; VON INS (2010) foi apresentado um acompanhamento do crescimento do número de publicações científicas no período de 1907 a 2007, os dados indicaram uma taxa de crescimento de cerca de 5,6% ao ano e um período de de 13 anos para duplicação desse número. Segundo os autores, o número de periódicos registrados em 1950 foi de cerca de 60.000 e a previsão para o ano 2000 foram cerca de 1.000.000 de publicações. Segundo a matéria do THE ECONOMIST (2010), a IDC (*International Data Corp*), empresa de pesquisa de mercado, em média 1.200 exabytes de dados digitais foram gerados apenas no ano de 2008.

Sob a óptica da produção acadêmica, onde a relevância e acessibilidade de uma informação tem grande peso, o volume de informações disponíveis no ciberespaço pode acarretar alguma dificuldade à pesquisadores, estudantes e profissionais, quando em desenvolvimento de um trabalho científico.

Segundo estimativa divulgada em matéria do SENADO (2010) publicações nacionais cresceram cerca de 572% no total mundial em 25 anos. Usando apenas esse exemplo é possível afirmar que o volume de dados para leitura e pesquisa é bastante significante, contudo nem todas as publicações são relevantes para

determinadas pesquisas acadêmicas. Sob essa perspectiva, a garimpagem e gerenciamento de trabalhos que sejam relevantes e também relacionados ao trabalho desenvolvido cabe ao esforço manual por parte do pesquisador.

A busca de informações demanda ainda mais tempo e atenção quando o acadêmico não tem conhecimento aprofundado quanto às publicações e autores relevantes na área a qual pretende abordar em seu trabalho. Segundo MANNING; RAGHAVAN (2009), 92% dos usuários da Internet afirmam que esta é a melhor fonte para busca de informações. É comum realizar pesquisas nas grandes ferramentas de busca, como o Google[1] e o Yahoo[2], que resultam em uma enorme lista de *links* melhores ranqueados que contenham as palavras inseridas pelo usuário, contudo a resposta à dúvida que iniciou a busca não está explícita. As figuras 1.1 e 1.2 ilustram uma solicitação de busca e um trecho do resultado obtido.

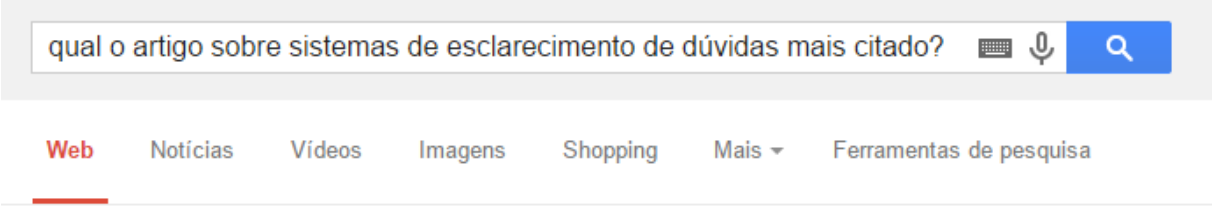

Figura 1.1: Exemplo de pesquisa no Google

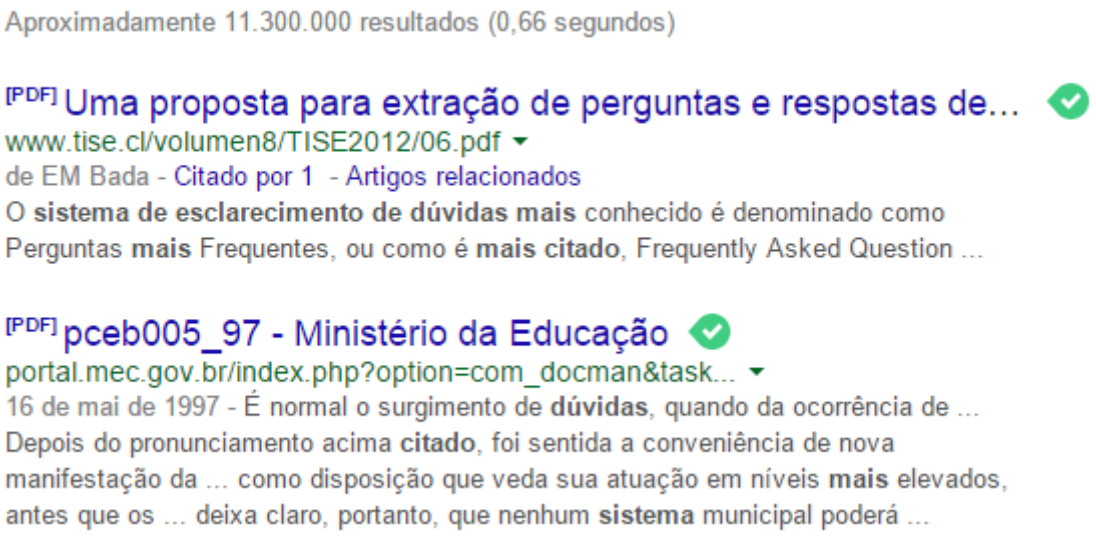

Figura 1.2: Recorte do exemplo de resultado da busca no Google

[1] http://www.google.com
[2] http://www.yahoo.com

Fazer perguntas é a forma mais natural de expressão do ser humano quando em busca de conhecimento. A linguagem, bem como a percepção, dedução, memória, raciocínio e cognição permitem uma aprendizagem além da experiência interativa. Expressar-se e ser compreendido em linguagem natural é um fator determinante para o processo de aprendizado.

Para encontrar respostas às muitas perguntas iniciadas a partir das primeiras incursões em sua pesquisa, o usuário deve realizar uma tarefa complexa, selecionar e ler diversos documentos que tenham alguma relação com o trabalho pretendido e, mesmo diante desse esforço, alguns dos documentos pré-selecionados podem não conter resposta satisfatória ao questionamento inicial.

Mesmo com todas as adaptações dos algoritmos responsáveis pela busca nos grandes portais de busca existentes, uma deficiência destes é ausência da capacidade de dedução (ZADEH, 2006). Sistemas de Perguntas e Respostas têm sido explorados no intuito de preencher essa lacuna.

Sistemas de Esclarecimento de Dúvidas, ou Sistemas de Perguntas e Respostas (*Question Answering Systems*), têm por objetivo recuperar informação precisa dentro de uma grande coleção de documentos (KANGAVARI; GHANDCHI; GOLPOUR, 2008). Sistemas Q&A provém de uma interseção de disciplinas, envolvendo Inteligência Artificial, Processamento de Linguagem Natural, Gerenciamento de Conhecimento e Dados além de Ciência Cognitiva (GUPTA; GUPTA, 2012).

Dentro do campo acadêmico, explora-se Sistemas de Perguntas e Respostas, ou Sistemas de Esclarecimento de Dúvidas, como meio de interação entre o aluno e um especialista ou a interação entre o aluno e o ambiente computacional, com o intuito de solucionar questionamentos do aluno ou até gerados pelo próprio especialista, esta forma de interação é de fundamental importância tanto para a consolidação do conhecimento, quanto a sua validação. O formato eletrônico diminui custos e facilita a replicação, armazenamento e compartilhamento de documentos, o que favorece a utilização da Web em processos de ensino à distância ou como apoio ao ensino convencional (SAIAS, 2010).

Esse tipo de sistemas buscam simular a naturalidade que ocorre em um diálogo entre duas pessoas, quando pensa-se em esclarecer uma dúvida, não é esperado que o respondedor apresente uma lista de documentos candidatos a conter a resposta.

Segundo HIRSCHMAN; GAIZAUSKAS (2001), para responder à uma pergunta, o sistema precisa analisar a questão, encontrar uma ou mais respostas consultando fontes online e só então apresentar ao usuário de forma apropriada. A resposta apresentada precisa conter contexto suficiente para que seja validada. Contudo, encontrar a resposta exata é um dos problemas mais importantes em Sistemas de Esclarecimento de Dúvidas (KANGAVARI; GHANDCHI; GOLPOUR, 2008).

1.1 Motivação

O desenvolvimento e produção de um trabalho acadêmico exige grande esforço, atenção e tempo do autor. Atualmente, existem ferramentas computacionais capazes de auxiliar o acadêmico em diversas etapas de organização e gerenciamento do projeto (elaboração de fichamentos, filtragem de artigos relevantes, gerenciar artigos e referências utilizadas, etc.), tais como o Mendeley[3], Zotero[4], StArt[5], entre outros. Contudo ainda é raro encontrar apoio à maioria das tarefas em um só ambiente de trabalho.

Quando iniciando a busca por um tema para início da produção acadêmica é natural que sejam levantados diversos questionamentos, acerca das publicações sobre o tema, sejam as mais recentes ou as mais citadas, os autores mais citados ou as publicações de um determinado autor sobre o tema a ser explorado. Questionamentos dessa natureza são passos necessários para apropriação do tema que se deseja pesquisar.

Ao realizar uma pesquisa desse gênero em quaisquer ferramentas de busca disponíveis online, os resultados são muito extensos, abrangendo páginas da web, imagens, vídeos, documentos, entre outros, que contenham palavras inseridas nas buscas. Cabe ao usuário filtrar as informações, de forma manual, e destacar as que julgar relevante à sua pesquisa. Essa grande quantidade de informação pode acabar se tornando um empecilho para essa busca direta por respostas explícitas.

[3] http://www.mendeley.com
[4] http://www.zotero.org
[5] http://lapes.dc.ufscar.br/tools/start_tool

Por esse motivo, foi identificado a necessidade do desenvolvimento de um ambiente inteligente e colaborativo capaz de acompanhar e apoiar a produção acadêmica, agregando apoio tanto de elementos computacionais quanto humanos (especialistas), capaz de identificar a pergunta inserida pelo usuário e fornecer resposta direta, além de informações de referência como documento de onde foi extraída a resposta e recomendações de outros artigos relacionados.

1.2 Objetivo

Conceber um Sistema Esclarecimento de Dúvidas, integrante do conjunto de funcionalidades que compõe o Ambiente Inteligente e Colaborativo de Apoio à Produção Acadêmica (AICAPA), capaz de identificar uma pergunta inserida em linguagem natural do tipo WH (o que, qual, quais, quando, quem) e apresentar respostas adequadas de forma automática ao usuário.

Para realizar esse objetivo geral, foi necessário estabelecer os seguintes objetivos específicos:

- Conceber uma arquitetura de Sistemas de Esclarecimento de dúvidas e suas interfaces, integrada ao conjunto de funcionalidades do AICAPA, estabelecendo as funcionalidades de cada módulo que o compõe e as interações necessárias;
- Aplicar técnicas de Processamento de Linguagem Natural para identificar o tipo de pergunta inserida pelo usuário e buscar melhores candidatos à resposta tanto na web quanto no banco de conhecimento do Ambiente;
- Desenvolver um protótipo como prova de conceito da metodologia aplicada.

1.3 Metodologia

Para composição deste trabalho foi realizada uma revisão bibliográfica acerca do tema abordado, priorizando trabalhos que apliquem técnicas de Processamento de Linguagem Natural. O desenvolvimento deste pode ser compreendido nas três etapas seguintes:

- Projeto: Definição do tema a ser abordado, levantamento de publicações relacionadas ao tema e revisão bibliográfica;

- Desenvolvimento: Concepção da modelagem da proposta;
- Prova de conceito: Implementação do protótipo e avaliação dos resultados.

1.4 Organização da Dissertação

O presente trabalho está organizado da seguinte forma:

O Capítulo 1 apresenta, de forma introdutória, o conteúdo deste trabalho. Destacando a motivação, os objetivos e a metodologia do desenvolvimento deste.

O Capítulo 2 apresenta o Ambiente do qual o subsistema descrito neste trabalho faz parte, abordando as principais funcionalidades, os módulos que o compõe e a arquitetura do sistema.

No Capítulo 3 é apresentado um breve histórico da evolução dos Sistemas de Esclarecimento de Dúvidas, algumas descrições destes sistemas, sua arquitetura, características e os principais problemas identificados nesta área. Além de uma breve revisão de outros trabalhos sobre este tema.

O Capítulo 4 apresenta o Agente de Dúvidas, a arquitetura proposta para esse subsistema, os módulos que o compõe e o fluxo de atividades.

No Capítulo 5 apresentamos a prova de conceito da proposta apresentada, os componentes tecnológicos utilizados na concepção e desenvolvimento do protótipo que valida a modelagem concebida, além da análise dos resultados obtidos.

O Capítulo 6 finaliza este trabalho com as considerações finais, as lições aprendidas, limitações do trabalho e as sugestões para trabalhos futuros.

2 CONTEXTO DO PROBLEMA – O AMBIENTE AICAPA

O AICAPA (Ambiente Inteligente e Colaborativo para Apoio à Produção Acadêmica) trata-se de um sistema que propõe-se a integrar conhecimento das áreas de recuperação de informação e recomendações para apoiar, computacionalmente, a produção acadêmica. O AICAPA foi concebido no Laboratório de Informática na Educação (LIEd) da Universidade Federal do Espírito Santo, que busca propor soluções computacionais de apoio à aprendizagem.

Dentre as etapas de diversas metodologias de pesquisa, o levantamento e análise da bibliografia disponível é um dos principais meios de apropriação sobre o tema explorado. A leitura e compreensão das publicações são os parâmetros para a identificação do problema e a definição da solução proposta.

Para o desenvolvimento destas atividades são utilizadas ferramentas que, além de auxiliar na busca das publicações, possa contribuir no gerenciamento das informações adquiridas ao longo do processo, contudo parte das atividades ainda é apoiada somente por esforço pessoal do acadêmico, como a etapa de fichamentos e seleção de artigos relevantes.

A proposta do ambiente colaborativo partiu da análise de ferramentas de gerenciamento, revisão bibliográfica e revisão sistemática disponíveis atualmente. Para essa etapa de análise foram escolhidas ferramentas bem difundidas e de grande aceitação no meio acadêmico, que foram: Mendeley, EndNote, PaperBox, Zotero e StArt.

Foi observado que, apesar da qualidade e diferentes funcionalidades oferecidas pelas ferramentas, estas têm foco em uma ou duas atividades de apoio à produção acadêmica, geralmente gerenciamento de referências bibliográficas, formatador de citações, gerenciamento de artigos e revisão sistemática.

Dentro das tarefas diretamente relacionadas à produção acadêmica, podemos também destacar a elaboração de fichamentos, acompanhamento da evolução, esclarecimento de dúvidas, recomendação de trabalhos relacionados, entre outras.

Diante das características analisadas nos softwares supracitados foi concebido um ambiente capaz de atender as principais atividades ligadas à produção acadêmica,

unindo conhecimentos das áreas de Recomendação, Recuperação e Extração de Informação e Esclarecimento de Dúvidas. A descrição das características das ferramentas analisadas e a tabela de resumo comparativo estão apresentadas na seção 2.3. Já a descrição do ambiente proposto foi abordada na seção 2.1.

2.1 Características do Ambiente

O principal objetivo do AICAPA é fornecer aos pesquisadores e estudantes um ambiente integrado, que apoie o processo de busca e seleção de artigos relevantes na web, que realize recomendações de artigos relevantes baseadas nos interesses do usuário, que seja capaz de identificar perguntas e dúvidas geradas ao longo do processo e poder apresentar resposta satisfatória de forma automática. Além de um ambiente que permita ao usuário a produção e gerenciamento dos fichamentos com maior facilidade.

Além de priorizar a automatização de tarefas repetitivas e que demandam grande esforço do acadêmico, não podemos ignorar a crescente difusão do trabalho colaborativo. Segundo DIAS (2001), os processos e estratégias colaborativas integram uma abordagem educacional na qual os alunos são encorajados a trabalhar cooperativamente na resolução de problemas.

Sob essa perspectiva, o sistema proposto proporciona um ambiente inteligente que promove o compartilhamento de conhecimento, promovendo interações entre usuários de diferentes níveis (novos usuários e mais experientes, doutores, mestres ou mestrandos e pesquisadores de todos os níveis, etc), apoiando o trabalho coletivo e a troca de experiências pessoais.

Além das funcionalidades técnicas propostas, o AICAPA pretende proporcionar interação social entre seus usuários. Das funcionalidades com aspectos sociais, podemos destacar: a recomendação direta de usuários, encaminhar perguntas para especialistas contidos no ambiente, os grupos de trabalho, além dos perfis pessoais.

As funcionalidades e características dos agentes que compõe o AICAPA são apresentados na seção 2.2.

2.2 Arquitetura Geral do AICAPA

A arquitetura proposta foi dividida em módulos, que representam as principais funcionalidades do sistema, que são: perfil do usuário, módulo de interface, módulo de recomendação, módulo de perguntas e respostas, além dos bancos de dados, de usuário e artigos. A figura 2.1 ilustra as interações entre eles, em destaque está o Agente de Dúvidas, que é o recorte deste ambiente que foi explorado na presente dissertação de mestrado.

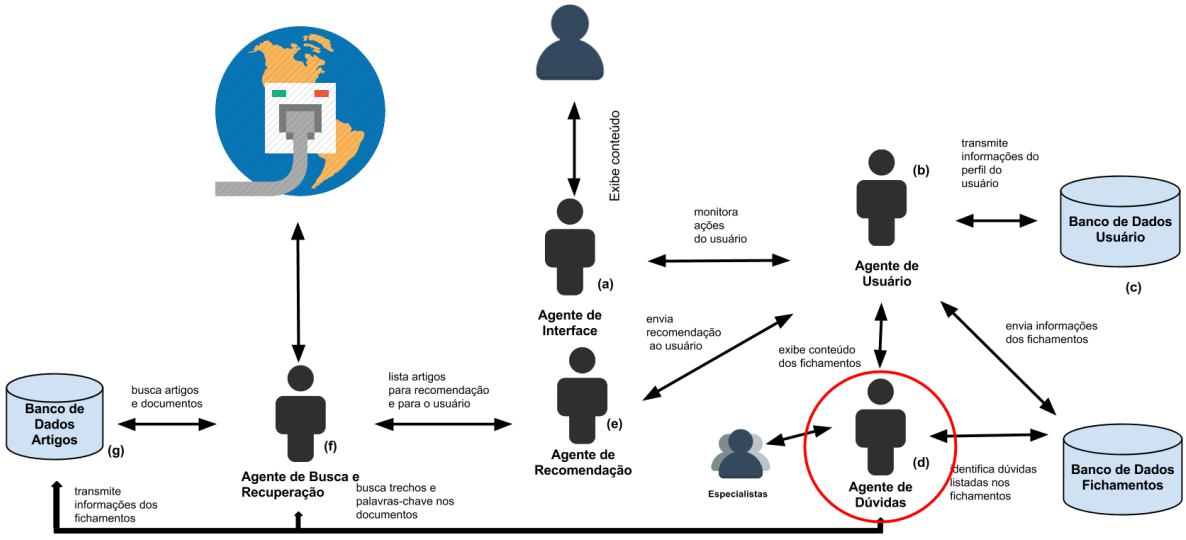

Figura 2.1: Arquitetura proposta - AICAPA

2.2.1 Agente de Interface

No Agente de Interface, identificado na figura 2.1 pela letra **(a)**, acontecem as interações entre o usuário e o sistema em si. Este componente é responsável por monitorar todas as ações realizadas no ambiente, as informações relacionadas à essa atividade serão armazenadas por esse agente, criando um histórico de atividades que será a referência para a atualização automática dos perfis pessoais. O Agente de Interface é também responsável por acionar os demais agentes.

2.2.2 Agente de Usuário

O Agente de Usuário **(b)** recebe os dados inseridos no sistema, que são então encaminhados ao Banco de Dados de Usuário **(c)**. A cada nova interação captada pelo agente de interface que contenha informação relevante (novos interesses, novos fichamentos, entre outras), os dados serão então repassados ao Agente de Usuário, que atualizará o perfil pessoal automaticamente.

O usuário, ao realizar o cadastro no ambiente, envia uma cópia do currículo, e o sistema então recupera as informações relevantes tais como: publicações do autor, nível de escolaridade, área de interesses. O perfil do usuário exerce papel fundamental neste ambiente. A partir das informações fornecidas no momento de cadastro, o AICAPA mapeia os interesses do usuário, que serão os parâmetros de definição para recomendações de artigos, grupos e do esclarecimento de dúvidas.

As informações do perfil do usuário também são parâmetros para a seleção de usuários especialistas. Essa categoria de usuários serão referência para determinadas ações do sistema, como a recomendação de usuários e a encaminhamento de dúvidas listadas nos fichamentos. A interação com o ambiente também é usada como critério de seleção.

2.2.3 Banco de Dados do Usuário

O componente de armazenamento de informações que abrange o perfil e interações do usuário com o sistema é o Banco de Dados do Usuário. Este componente é atualizado sempre que o usuário executa alguma ação no AICAPA. Também recebe informações dos outros Agentes, como o Agente de Recomendação e Agente de Dúvidas, o primeiro envia informações sobre os interesses adquiridos pelo usuário, enquanto o segundo envia informações a respeito das perguntas que foram enviadas pelo usuário e as respostas que ele pode ter enviado.

2.2.4 Banco de Dados de Artigos

O Banco de Dados de Artigos **(g)** armazena todos os dados relacionadas aos artigos: informações do cabeçalho, autor(es), ano de publicação, resumo, além das referências listadas pelo autor(es) do artigo. O objetivo para esse tipo de armazenamento de dados é poder realizar buscas de trabalhos relacionados dentro das referências de cada artigo do banco. A recuperação e, consequente, apresentação ao usuário pode auxiliá-lo na seleção manual de referências.

2.2.5 Agente de Recomendação

O Agente de Recomendação **(e)** é responsável por sugerir documentos, pessoas e grupos semelhantes ao(s) perfil(s) do(s) usuário(s). O agente recebe os dados do perfil, enviados pelo Agente de Usuário, realiza cálculos de similaridade e localiza documentos, pessoas e grupos adequados àquele usuário.

Este agente também interage com o Agente de Dúvidas, fornecendo informações relevantes às dúvidas listadas pelo usuário. Pode, também, sugerir outras perguntas previamente respondidas e que tenham relação com os questionamentos enviados.

2.2.6 Agente de Busca e Recuperação

O Agente de Busca e Recuperação **(f)** atua diretamente com os Agentes de Recomendação e Dúvidas, é responsável por recuperar informações sobre os documentos armazenados no Banco de Dados de Artigos, além de enviar informações para este componente. Em ocasiões de ausência de informação no Banco de Dados, este agente realiza buscas diretas na Internet e apresenta os resultados ao agente que requisitou a busca.

2.2.7 Agente de Dúvidas

Já o Agente de Dúvidas **(d)** tem por objetivo fornecer respostas de forma automática às perguntas enviadas pelos usuários em linguagem natural. Depois que as respostas são apresentadas ao usuário, o agente envia o par pergunta-resposta ao seu banco de conhecimento, para que seja consultado inicialmente no intuito de fornecer respostas a perguntas similares. As perguntas que não forem respondidas de forma adequada são enviadas aos usuários especialistas, determinados pelo Agente de Usuário com base nas informações recuperadas do perfil dos colaboradores do ambiente.

Este agente ainda interage com os componentes de fichamentos, que permitem ao usuário listarem dúvidas ocasionadas durante e após a leitura dos artigos selecionados. O agente então aciona os Agentes de Recomendação e Busca, que recuperam informações relevantes às dúvidas e as apresentam ao usuário. Os questionamentos listados nos fichamentos podem, ainda, ser encaminhados aos usuários especialistas ou aos grupos de discussão.

O Agente de Dúvidas é a porção do AICAPA explorado neste trabalho e, por isso, suas especificidades são apresentadas no capítulo 4.

2.2.8 Fichamentos

O AICAPA oferece ao usuário a possibilidade de produzir e gerenciar fichamentos das obras lidas. O ambiente dispõe de um componente de fichamento que apresenta estruturas textuais determinadas de acordo com o tipo de fichamento pretendido. As interfaces desse componente ainda permitem que o usuário localize o fichamento com maior facilidade, com filtros por datas ou título.

Em um dos modelos de fichamento disponibilizados pelo AICAPA, o usuário pode listar dúvidas a respeito do artigo lido. Essas dúvidas serão, então, identificadas pelo Agente de Dúvidas, que pode procurar por candidatos a resposta na base de conhecimento local ou enviar os questionamentos aos usuários especialistas determinados pelo AICAPA.

O AICAPA ainda permite que o usuário crie novas estruturas de fichamento, baseado nas necessidades identificadas ao decorrer da pesquisa deste acadêmico.

De acordo com LAKATOS; MARCONI (2003), a pesquisa bibliográfica de um trabalho científico acadêmico é dividida em oito (08) fases distintas, sendo uma delas a fase de fichamentos. As fichas, ou fichamentos, auxiliam o pesquisador na organização do material de pesquisa, armazenamento de citações, elaborar críticas, além de possibilitar ao acadêmico a assimilação do conteúdo. As fichas registram tudo que possa servir de embasamento para a pesquisa (MEDEIROS, 2006). Independente da sua finalidade, as fichas possuem três (03) componentes estruturais principais: cabeçalho, referência bibliográfica e corpo ou texto.

O cabeçalho reúne informações que permitem ao pesquisador identificar as obras que, possivelmente, vão compor a base referencial teórica do trabalho a ser desenvolvido. Segundo MEDEIROS (2006), o cabeçalho engloba o título genérico ou específico e a letra indicativa da sequência de fichas, um exemplo desse tipo de ficha é ilustrado pela figura 2.2.

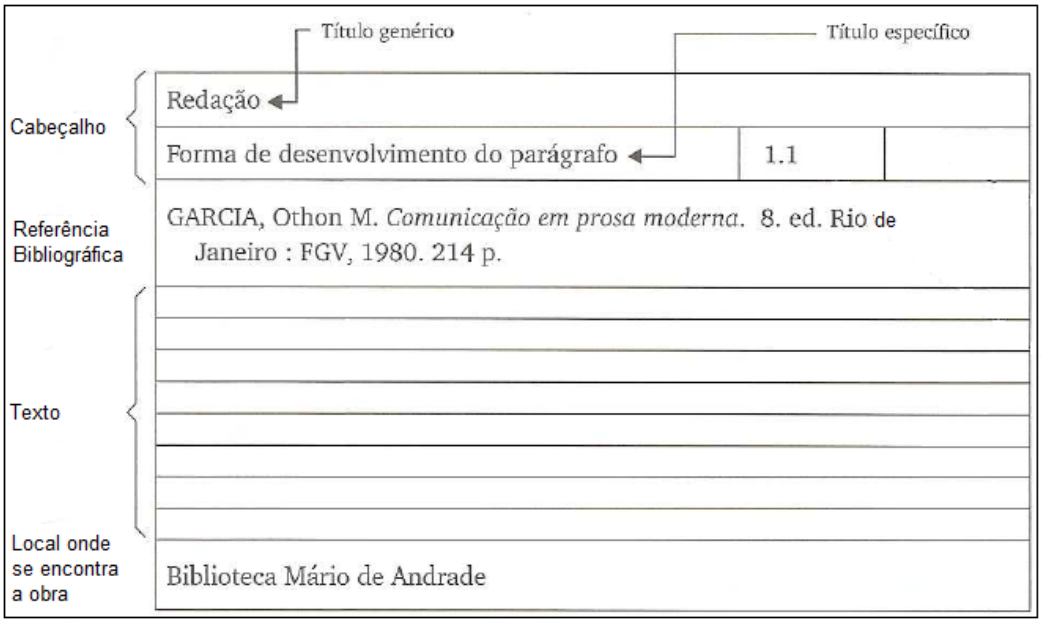

Figura 2.2: Elementos estruturais da ficha, retirado de Medeiros (2006)

Sob a perspectiva tecnológica, os fichamentos passaram a ter outra estrutura, variando de acordo com a necessidade do autor. Elementos estruturais como a localização da obra puderam ser substituídos pelo endereço eletrônico onde a publicação foi disponibilizada. A figura 2.3 ilustra um exemplo de fichamento elaborado a partir das necessidades identificadas na metodologia aplicada nesta pesquisa.

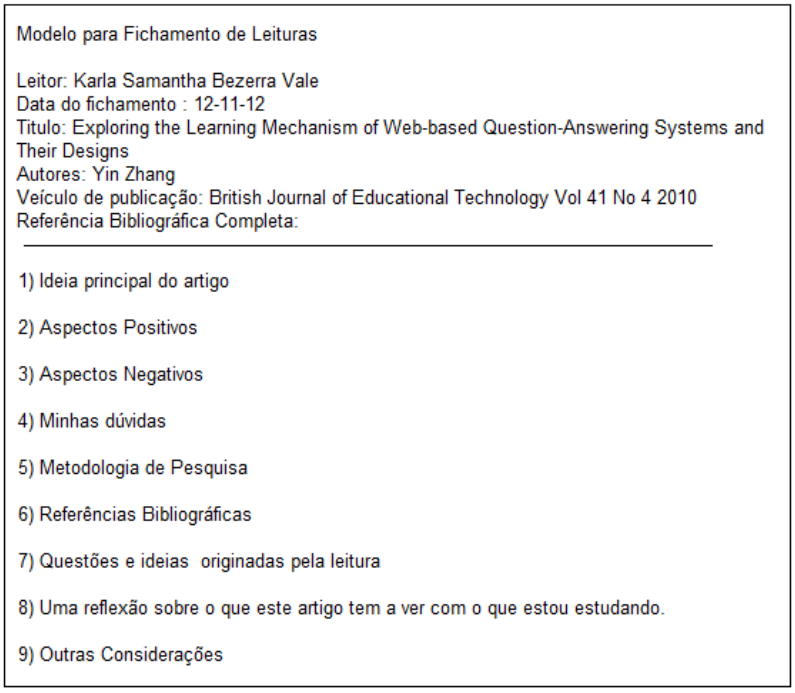

Figura 2.3: Elementos estruturais - exemplo de fichamento eletrônico

Tipos de Fichamento

Existem três (03) tipos básicos de fichamento: bibliográfico, citação e resumo ou conteúdo (KAUARK; MANHÃES; MEDEIROS, 2010). Fichamentos bibliográficos contém descrições dos tópicos abordados no texto lido, podendo conter comentários do autor.

Já os fichamentos de resumo ou conteúdo sintetizam as ideias apresentadas no texto. Esse tipo de fichamento ainda pode ser dividido em duas categorias: (a) Informativo: contém informações específicas, podendo ser objetivos, métodos, resultados e conclusões; (b) Indicativo: engloba descrições gerais sobre a obra analisada. Fichamentos de citação consistem na reprodução fiel de frases ou sentenças consideradas relevantes ao estudo em pauta (LAKATOS; MARCONI, 2003).

2.3 Ambientes de Apoio à Produção Acadêmica

Nesta seção são analisados os softwares que serviram de base para a modelagem e desenvolvimento do AICAPA.

Dentro da pesquisa acadêmica existem algumas ferramentas que fornecem ambientes que apoiam às atividades que compõe a concepção e elaboração do produto final, contudo esses ambientes ainda apresentam algumas deficiências, tais como a ausência de modelos para fichamentos, acompanhamento colaborativo, recomendação e esclarecimento de dúvidas.

Foram analisadas ferramentas que oferecem apoio à alguma das etapas de desenvolvimento do trabalho acadêmico, dentre as opções disponíveis atualmente, as selecionadas foram: Mendeley, EndNote, PaperBox, Zotero e StArt.

Uma das ferramentas de maior destaque nesta área é o Mendeley, desenvolvida por Glyph & Cog, é um gerenciador de referências bibliográficas, além de uma rede social acadêmica. Um de seus diferenciais mais marcantes é a possibilidade de criar grupos de discussão e, com isso, potencializar a colaboração entre diversos outros pesquisadores interessados em áreas correlatas.

Na ferramenta, o usuário pode visualizar o artigo, fazer anotações, destacar trechos e compartilhar o(s) artigo(s) com demais usuários. Nos grupos de discussão, os participantes podem disponibilizar artigos apenas para os membros, além de criar linhas de discussão no ambiente do grupo, semelhante ao padrão seguido pelas redes sociais atuais.

O Mendeley também conta com um sistema de busca, que permite localizar artigos e adicioná-los diretamente na biblioteca particular do usuário. Permite que o usuário sincronize a sua biblioteca local com a nuvem, podendo acessá-la em qualquer dispositivo com conexão à internet.

A figura 2.4 ilustra o detalhamento das informações de uma determinada publicação disponível na biblioteca particular do usuário.

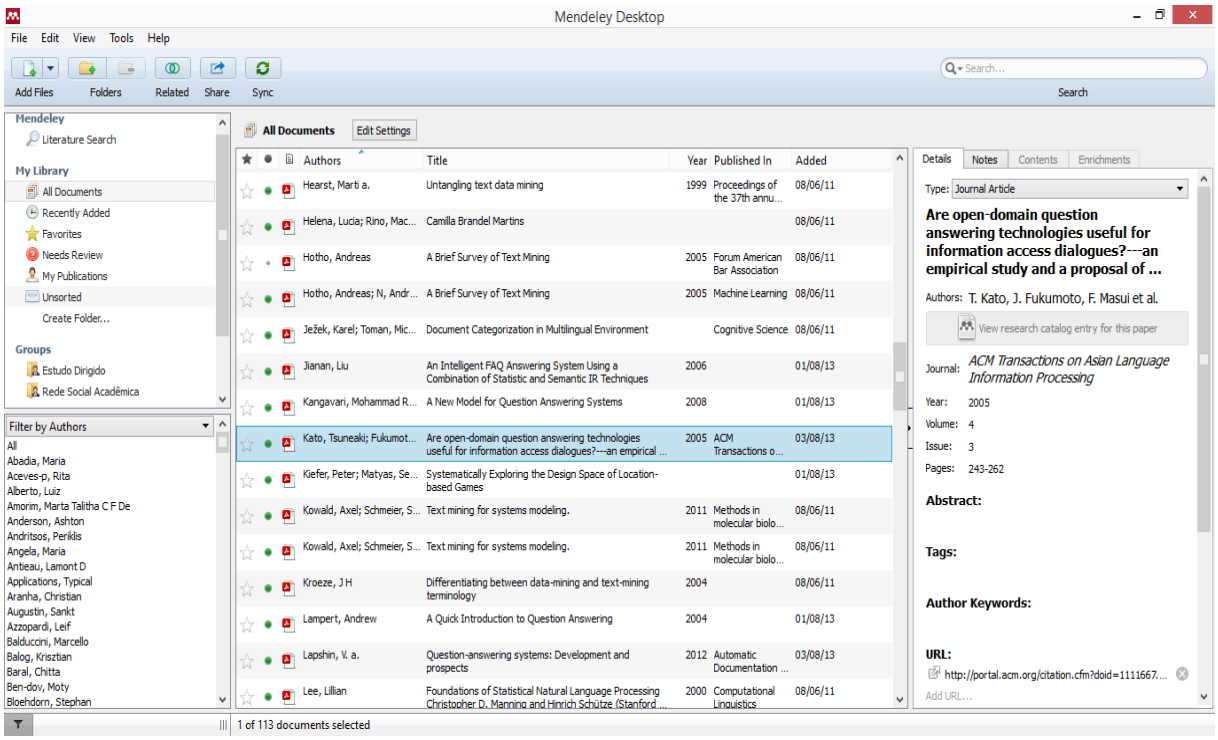

Figura 2.4: Interface Mendeley

O EndNote é uma ferramenta online desenvolvida pela Thomson Reuters para gerenciamento de referências, utilizado para organizar a bibliografia utilizada ao escrever trabalhos acadêmicos. Com esse software, o usuário pode criar um banco de dados de referências personalizado, organizar e armazenar citações em

bibliotecas. É, também, um formatador de citações e buscador, oferecendo acesso aos catálogos online diretamente de sua interface.

Na figura 2.5 está ilustrada o detalhamento da função de edição das informações acerca da publicação selecionada.

Uma das desvantagens do EndNote é o fator financeiro, diante das outras ferramentas esta é a única que apresenta menor custo benefício. Sendo que os usuários têm, apenas, uma licença temporária (30 dias) para uso gratuito.

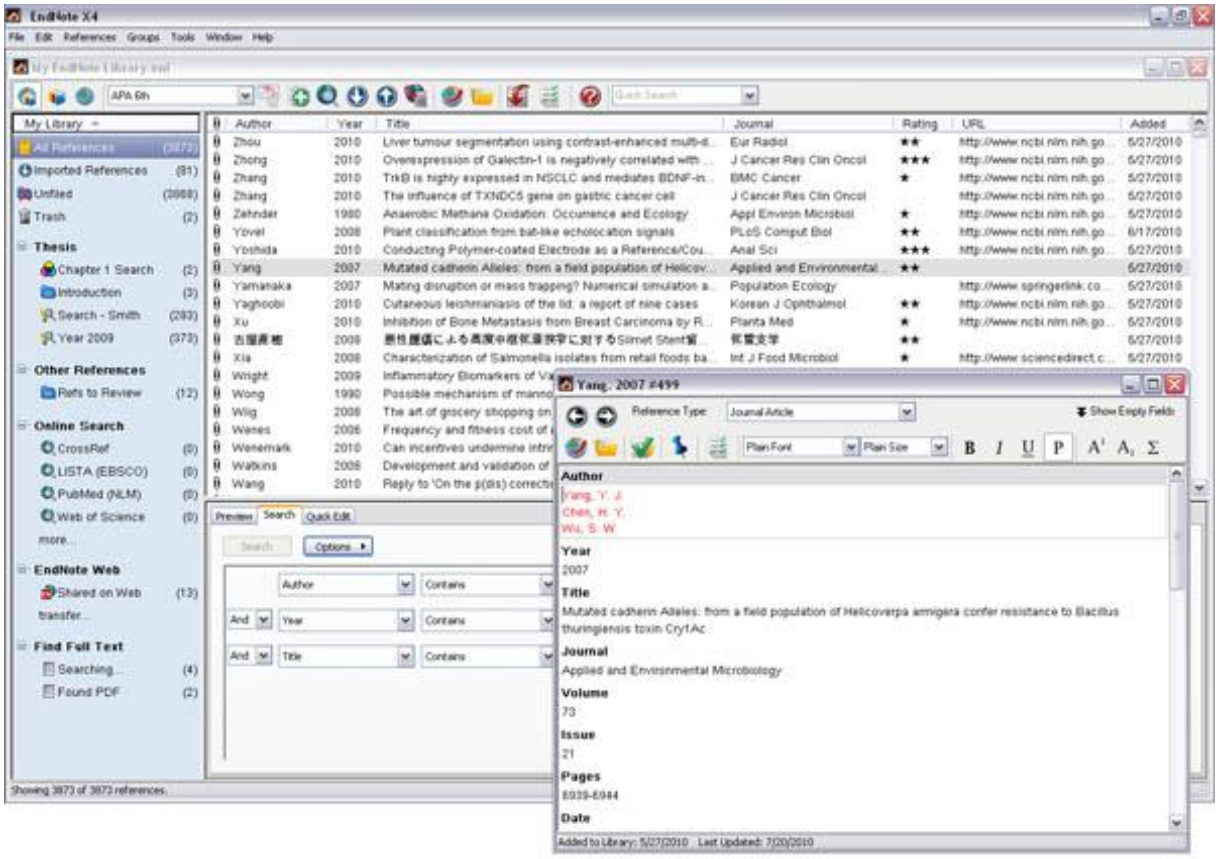

Figura 2.5: Interface EndNote

PaperBox é uma ferramenta baseada em nuvem, que permite a organização, formatação e o acesso das referências bibliográficas de qualquer dispositivo com acesso à internet. Esta ferramenta permite o compartilhamento de informações entre usuários, além de oferecer um *plug-in* para o pacote Office.

A figura 2.6 ilustra a expansão das informações acerca da publicação selecionada na biblioteca curada pelo usuário. Nesse detalhe o usuário tem, à sua

disposição, a possibilidade de editar as informações, criar linhas de discussão sobre a publicação e realizar buscas acerca desta publicação.

A ferramenta também permite que usuários compartilhem arquivos de texto, criar alertas sobre novas publicações - estas serão acrescidas na biblioteca particular do usuário de forma automática, exportar a sua biblioteca particular e sincronizar a sua conta em todos os dispositivos vinculados à ela.

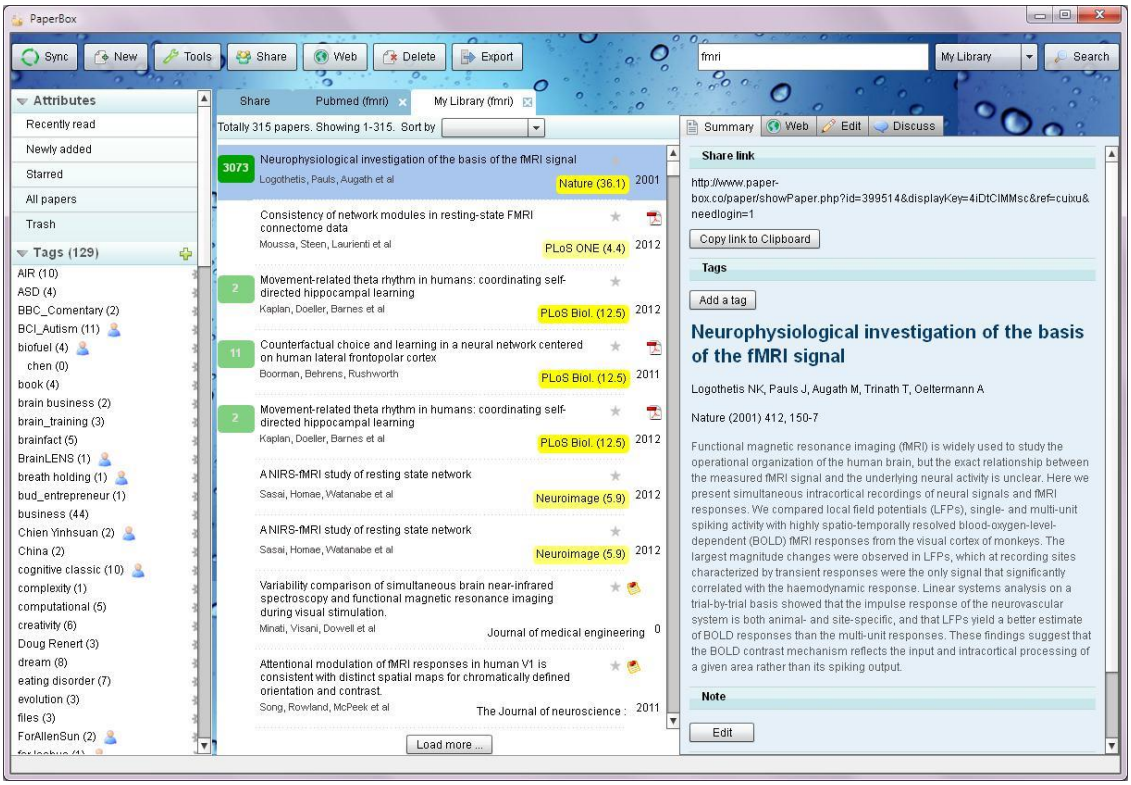

Figura 2.6: Interface PaperBox

Já o Zotero é uma ferramenta gratuita, desenvolvida pelo *Roy Rosenzwieg Center for History and New Media*, que oferece ao usuário as funcionalidades de organizar, colecionar e compartilhar arquivos de referências bibliográficas. É possível adicionar notas, marcações aos itens listados no arquivo. Também é possível anexar artigos relacionados, de forma manual.

A ferramenta conta com um mecanismo de reconhecimento do conteúdo do navegador do usuário e a indexação automática dos arquivos enviados pelo usuário, facilitando a busca de trabalhos relevantes. Assim como nas demais ferramentas, o Zotero disponibiliza ao usuário a possibilidade de edição das informações acerca da publicação, como visto na figura 2.7. O Zotero também disponibiliza um *plugin*

compatível com o Microsoft Word, que permite ao usuário criar citações de forma automática.

O StArt é uma ferramenta de apoio à revisão sistemática desenvolvida pelo Laboratório de Pesquisa em Engenharia de Software da UFScar (Universidade Federal de São Carlos). Seu objetivo é dar suporte ao planejamento, execução e análise final de uma revisão sistemática, independente do assunto ou área de pesquisa, tornando-a mais ágil, precisa e replicável.

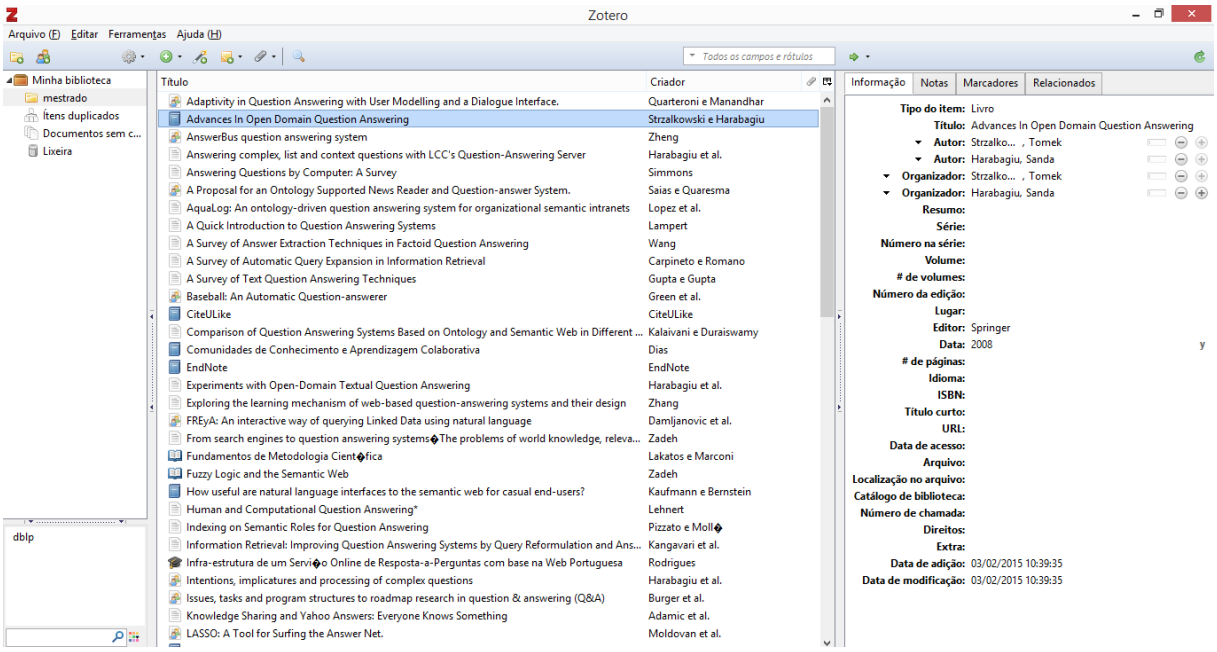

Figura 2.7: Interface Zotero

Para iniciar o processo de revisão, é necessário que o usuário preencha os critérios para a seleção e filtragem dos artigos. O usuário pode criar rotinas de busca, em portais como o Google Acadêmico e o IEEE, para isso a sentença (*string*) de busca precisa ser bem definida, caso contrário a ferramenta não apresenta resultados.

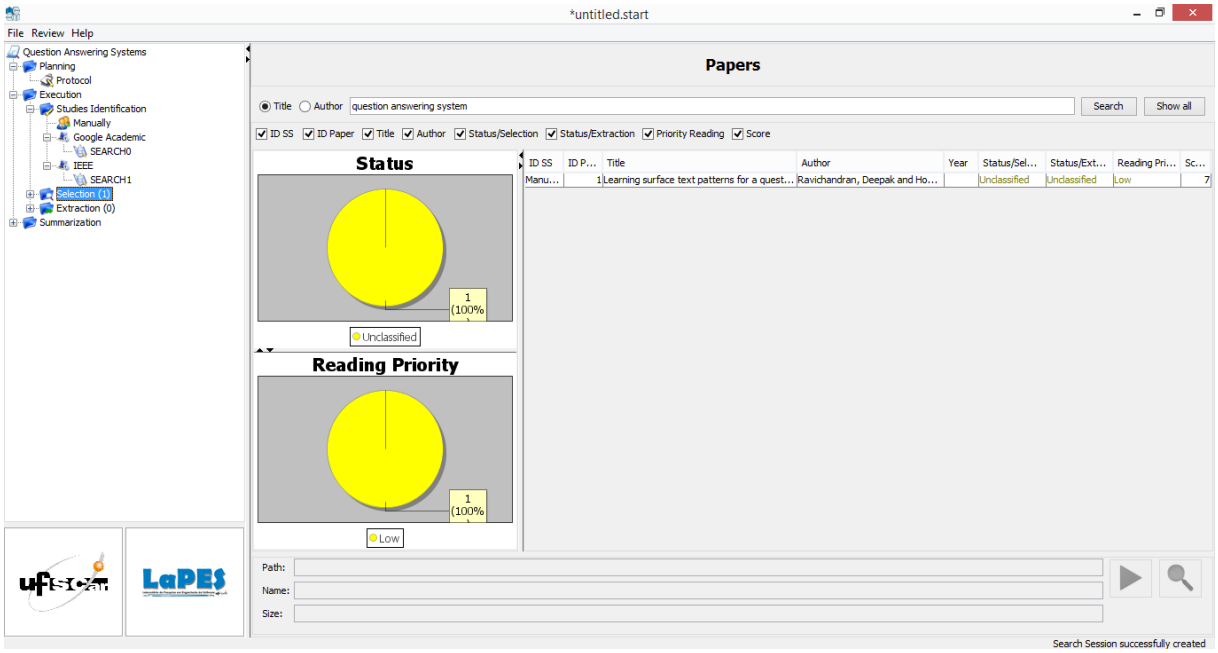

Figura 2.8: Interface StArt

Já a figura 2.8 ilustra os gráficos de seleção de artigos para leitura e priorização das publicações selecionadas.

2.3.1 Tabela Comparativa

Nesta seção estão representados em tabelas os resumos comparativos entre as ferramentas comerciais e as publicações que serviram como base para modelagem e implementação do ambiente colaborativo.

As ferramentas apresentam funcionalidades semelhantes, como gerenciamento de referências, edição de informações das publicações, anotações, entre outras, e também algumas particularidades, principalmente no caso da ferramenta StArt, que oferece apoio à revisão sistemática. Contudo, as características das ferramentas ainda não contemplaram as atividades de recomendação e esclarecimento de dúvidas, que são as principais funcionalidades disponibilizadas pelo AICAPA.

A tabela 2.1 apresenta, de forma comparativa, as características das ferramentas estudadas como referência para o AICAPA. Os critérios para análise foram: disponibilidade web, custo, realizar recomendações, busca e recuperação automatizada, solucionar dúvidas, apoiar a revisão literária e disponibilidade de armazenamento em nuvem.

Tabela 2.1: Resumo Comparativo

Softwares / Critérios	Mendeley	EndNote	PaperBox	Zotero	StArt
Disponibilidade web	Sim	Sim	Sim	Sim	Não
Custo	Gratuito	$250	Gratuito	Gratuito	Gratuito
Realiza Recomendações?	Não	Não	Não	Não	Não
Busca e Recuperação de Informação?	Não	Não	Não	Não	Sim
Esclarece dúvidas?	Não	Não	Não	Não	Não
Apoia a Revisão da Literatura?	Não	Não	Não	Não	Sim
Baseado em nuvem?	Sim	Sim	Sim	Sim	Não

2.4 Considerações finais do Capítulo

Este capítulo focou nas características e funcionalidades do AICAPA, a proposta de arquitetura e as atividades atribuídas a cada componente. Apresentamos, também, alguns dos ambientes de apoio à produção acadêmica que serviram de referência para a concepção e modelagem do AICAPA.

O ambiente AICAPA difere-se das ferramentas analisadas por, além de oferecer um ambiente colaborativo, automatizar atividades que tomam tempo e esforço intelectual por parte do acadêmico. Ao disponibilizar mecanismos automáticos de busca, recomendação e esclarecimento de dúvidas pretende-se poupar o acadêmico da tarefa de navegar inúmeras páginas em busca de informação relevante à sua pesquisa.

3 SISTEMAS DE ESCLARECIMENTO DE DÚVIDAS

Este capítulo descreve de forma aprofundada a base teórica dos componentes científicos explorados no desenvolvimento do protótipo apresentado nesta dissertação de mestrado. Aborda os conceitos de Processamento de Linguagem Natural contidos em Sistemas de Esclarecimento de Dúvidas.

A divisão deste capítulo segue a seguinte ordem: a seção 3.1 apresenta um breve histórico sobre os primeiro Sistemas de Esclarecimento de Dúvidas e a sua evolução, além das diversas conceituações e classificações desses sistemas, as características que definem um sistema desse tipo, apresenta uma visão geral da arquitetura desses sistemas e os problemas intrínsecos à eles.

3.1 Histórico dos Sistemas de Esclarecimento de Dúvidas

O interesse pela área de perguntas e respostas não é recente, uma das primeiras referências ao tema data de 1965, com o trabalho de SIMMONS (1965) que analisou quinze (15) diferentes sistemas, os sistemas listados incluem arquiteturas de sistemas de perguntas e respostas, interfaces de repositórios de dados, além de sistemas que buscam respostas para perguntas de fontes textuais (HIRSCHMAN; GAIZAUSKAS, 2001).

Um dos primeiros sistemas de perguntas e respostas foi o BASEBALL (GREEN JR. et al., 1961), que oferecia respostas acerca de uma determinada temporada da liga americana de baseball. Esse sistema analisava a pergunta, fazendo uso de conhecimento linguístico e, a partir disso, construía uma consulta (query) para uma base de dados. Dada a complexidade das consultas geradas, qualquer diferença entre o conceito da pergunta por parte do usuário e do sistema, poderia acarretar erros na obtenção da resposta.

Esse tipo de sistema, onde são oferecidas interfaces ou *front-end* para base de dados estruturadas, preocupavam-se com os seguintes tópicos (WEBBER; WEBB, 2010):

- Mapear questões dos usuários para consultas computáveis;

- Lidar com diferenças de conceptualização entre usuários e sistema;
- No caso de bases distribuídas, identificar de onde a informação necessária para construir a resposta deve ser importada.

Contudo, as primeiras incursões nesse tipo de sistemas Q&A foram abandonadas no final dos anos 80, por duas razões - uma social e uma técnica (WEBBER; WEBB, 2010). A técnica deu-se por conta do grande esforço necessário para garantir a efetividade e confiabilidade do mapeamento entre as queries dos usuários e do sistema. Quaisquer mínimas diferenças entre as perguntas deviam ser mapeadas em bases completamente diferentes e, em outros casos, o mapeamento depende de uma base específica. A dificuldade social estava ligada a ausência de usuários intermediários para esse tipo de tecnologia, já que usuários comuns (usuários sem conhecimento técnico), não tinham acesso às bases e gerenciadores de grandes bases de dados não tinham interesse em acessá-las.

Com o advento da internet, o problema social foi resolvido. Trazendo consigo as técnicas de aprendizado técnicas, que provaram ser úteis em outras áreas de tecnologia de linguagem, passando a ser aplicadas para problemas como mapeamento complexo entre queries de usuários e base de dados (WEBBER; WEBB, 2010).

Diversas conferências e workshops focaram nos aspectos ligados à área de perguntas e respostas (HIRSCHMAN; GAIZAUSKAS, 2001). Em 1999, a Conferência de Recuperação de Texto (TREC) patrocinou uma linha de pesquisa baseado em sistemas de perguntas e respostas, onde avaliava sistemas que ofereciam respostas a perguntas factuais consultando a base de documentos da TREC (corpus).

Em 2000, CARBONELL et al. (2000), delinearam uma visão ambiciosa para a pesquisa nessa área. A declaração indica um vasto espectro de perguntadores e a extensão de respostas baseado na conferência TREC-8, como visto na figura 3.1.

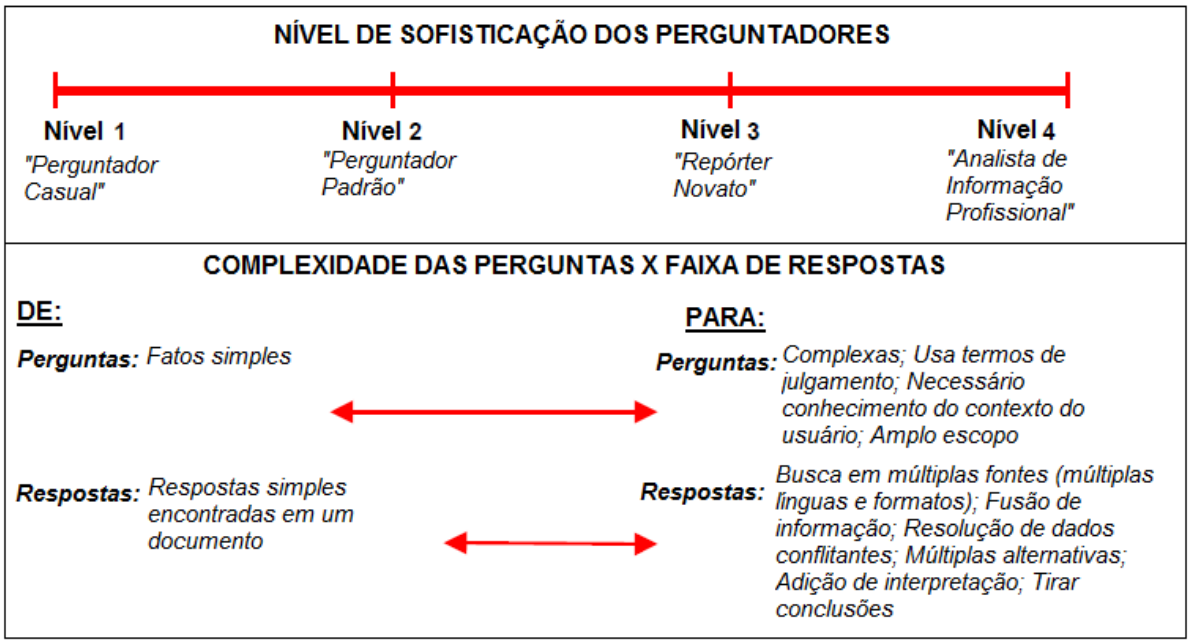

Figura 3.1: Relação Usuários x Perguntas e Respostas, adaptado de Carbonell et. al. (2000)

Mesmo com as contribuições e evoluções na pesquisa da área de Sistemas de Perguntas e Respostas, ainda existe muito campo a ser explorado. HIRSCHMAN; GAIZAUSKAS (2001) apontam a necessidade de sistemas que explorem as aplicações nas áreas de educação, tanto no processo de aprendizado quanto de propagação de conhecimento. Técnicas de avaliação automáticas podem ser utilizadas como meio de avaliação e classificação de notas, por exemplo. Os autores também levantam a necessidade de busca por respostas em diversas fontes de dados, usando outros formatos além das bases textuais.

Avanços nessa área ainda imprimem desafios para a implementação de sistemas, estas dificuldades são discutidas na próxima seção, bem como as características dos componentes comuns desses sistemas.

3.1.1 Classificação dos Sistemas

Em sistemas de recuperação de documentos tradicionais, a tarefa de extrair a resposta dos documentos recuperados é, inteiramente, do usuário, o que acaba tornando uma árdua tarefa cognitiva. Sistemas de perguntas e respostas tem o foco em reduzir essa tarefa, dando sequência à atividade de recuperação de documentos seguida de uma série de processamentos avançados com o intuito de localizar e apresentar a resposta (STRZALKOWSKI; HARABAGIU, 2008). Segundo MAYBURY (2004), um sistema de perguntas e respostas é um processo interativo humano-

computador que engloba uma compreensão precisa de informações dos usuários, tipicamente expressa em uma consulta em linguagem natural, recuperação de documentos relevantes, dados ou conhecimento a partir de fontes selecionadas, extraindo, qualificando e priorizando as respostas disponíveis dessas fontes e apresentando e explicando respostas de uma maneira eficiente.

Sistemas de esclarecimento de dúvidas podem ser classificados de diversas formas, sendo baseado na abordagem adotada, nas características de cada sistema (nível de compreensão e raciocínio, tipo de dados, base de conhecimento, generalidade), entre outras. MAYBURY (2003) classificou os sistemas baseando-se nas características de cada um, como visto na figura 3.2.

Segundo o autor, os Sistemas QA podem ser classificados de acordo com as características que o definem, mas não necessariamente estão restritos à elas. Como observado na figura 3.2, todos os sistemas listados na primeira coluna podem ser classificados nas relações estabelecidas entre as características que correspondem às perguntas e sua complexidade, volume da fonte e qualidade desta, corpus (base textual) e integração e geração da resposta.

Tipos de QAS	Pergunta / Complexidade da resposta	Volume da fonte / Qualidade	Corpus / Modelo de fonte	Integração e geração da resposta	Tipos de query do usuário	Tipo de resposta do sistema
TREC QAS Dicionários online, enciclopédias	Moderada / Fácil	Pequena (100 mb), estático / Alta qualidade	Enciclopédia Manuais técnicos	Fácil	FORMA: - Palavras-chave - Frases - Perguntas TIPO: - Quem - O que - Quando - Onde - Como - Por que - E se INTENÇÃO: - Requisição - Comando - Informar	- Entidades nomeadas - Frases - Factóide - Link para documento - Sumário
Manuais online Web QAS	Fácil a moderada / Moderada	Pequena a grande (10 GB), dinâmico / Fontes de qualidade variáveis	Web	Moderada		
QAS Multimídia, multilingual	Difícil / Difícil (interage com multilinguagem)	Muito grande, resposta em tempo real, dinâmico / Fontes de qualidade variáveis	Variada Multilinguagem	Difícil		

Figura 3.2: Tipos de Sistemas QA, adaptado de Maybury (2003)

MOLDOVAN et al. (2003) propuseram uma taxonomia baseada em critérios que desempenham papel importante na construção de um Sistema de Perguntas e Respostas:

- **Classe 1** - Sistemas capazes de processar perguntas factuais: Esses sistemas extraem respostas como trechos de texto de um ou mais documentos. A resposta é, frequentemente, encontrada em um texto ou com variações morfológicas simples. Tipicamente, as respostas são extraídas usando métodos empíricos baseados em palavras-chave.
- **Classe 2** - Sistemas habilitados com mecanismos simples de raciocínio: A característica dessa classe é o uso de inferências para relacionar a pergunta à resposta. São necessários métodos mais elaborados de detecção da resposta como ontologias ou codificação de conhecimento pragmático. Alternações semânticas, axiomas de conhecimento comum e métodos de raciocínio também são necessários.
- **Classe 3** - Sistemas capazes de fundir respostas de diferentes documentos: Nessa classe, partes da informação sobre a resposta estão espalhadas entre vários documentos, a construção da resposta faz-se necessária. A complexidade varia entre a montagem de listas simples até perguntas mais complexas, como perguntas do tipo script e template.
- **Classe 4** - Sistemas interativos: Esses sistemas são capazes de responder perguntas no contexto de interações com o usuário. Conforme mencionado por HARABAGIU et al. (2001), processar uma lista de perguntas dentro de um contexto envolve resolução de referência complexo.
- **Classe 5** - Sistemas capazes de raciocínio analógico: A característica desse tipo de sistemas é a habilidade de responder perguntas especulativas, como por exemplo: *"Como está a economia Brasileira?"*. Como a resposta para esse tipo de pergunta provavelmente não está explícita em algum documento, sistemas dessa classe decompõe a pergunta em queries que extraem porções de evidências, em seguida a resposta é formulada usando raciocínio por analogia.

GUPTA; GUPTA (2012) dividiram esses sistemas em dois grandes grupos: o primeiro grupo são os sistemas que adotam técnicas simples de processamento de linguagem natural e métodos de recuperação de informação, já o segundo engloba os

sistemas que são dependentes de raciocínio sobre a linguagem. Os dois grupos foram comparados de acordo com as características das dimensões de cada um, como a técnica adotada, o tipo de pergunta, domínio, entre outras. A tabela 3.1 apresenta a comparação feita pelos autores.

Tabela 3.1: Classificação proposta por Gupta e Gupta (2012)

Dimensões	**QAS baseados em PLN e RI**	**QAS baseados em raciocínio com PLN**
Técnica	Processamento sintaxe, Marcação de entidade nomeada e Recuperação de Informação	Análise Semântica ou raciocínio profundo
Fonte de dados	Documentos de texto	Base de conhecimento
Domínio	Independente de domínio	Orientado a domínio
Respostas	Extração de trechos de texto	Respostas sintetizadas
Tipos de pergunta	Perguntas do tipo WH	Perguntas WH, listas, dedutivas
Avaliação	Técnicas de recuperação de informação	-----

Já ZHENG (2002) classificou os sistemas a partir do domínio, aberto e fechado. No domínio fechado, a base de conhecimento já é estabelecida. Em sistemas de domínio aberto, a fonte de dados para busca e extração de respostas é a web. Apesar de sistemas de domínio aberto sejam, potencialmente, mais úteis que sistemas de domínio específico (fechado), a construção de sistemas deste tipo implica em mais problemas técnicos e teóricos, como afirmam DAMLJANOVIC; AGATONOVIC; CUNNINGHAM (2012). Contudo, (KAUFMANN; BERNSTEIN, 2007) afirmam que quanto mais um sistema é customizado à um domínio, melhor é a sua performance de recuperação de resposta.

A seguir são apresentados os componentes básicos de um Sistema de Esclarecimento de Dúvidas e suas respectivas funcionalidades.

3.1.2 Arquitetura Geral dos Sistemas QA

De acordo com WEBBER; WEBB (2010), um sistema básico de perguntas e respostas envolve uma cascata de processos, que tem por entrada a pergunta do usuário e a saída é uma lista de candidatos a resposta, de forma ranqueada, com indicação da fonte da informação apresentada. Essa cascata é, usualmente, dividida em quatro (04) módulos, que tem funções e atividades bem estabelecidas: análise da pergunta, seleção e extração da resposta, base de conhecimento e geração da resposta. Um exemplo desse tipo de arquitetura está representado na figura 3.3.

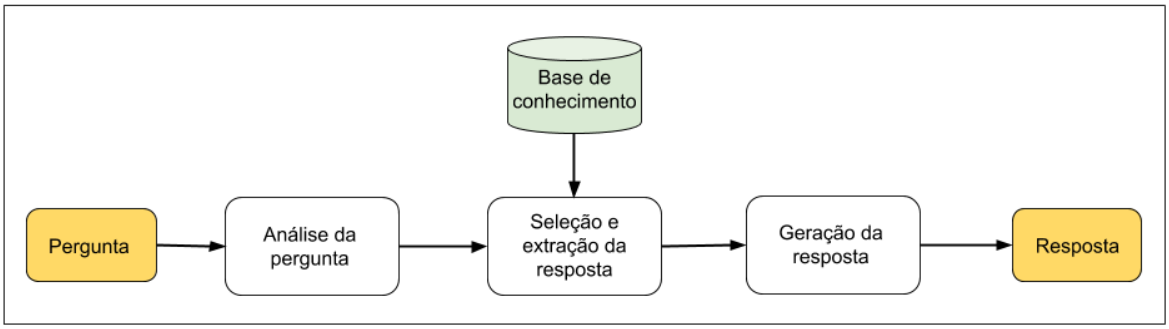

Figura 3.3: Arquitetura genérica de um sistema Q&A, adaptado de Maybury (2004)

Para buscar a resposta adequada é necessário saber o que procurar. O papel do componente de análise da pergunta, às vezes processamento da pergunta, é determinar o tipo da pergunta e o tipo de resposta esperado, construir o foco para a resposta e transformar a entrada em consultas para a ferramenta de busca (MOLDOVAN et al., 1999). A identificação do tipo da pergunta visa associar uma marcação, determinando o tipo de informação a ser buscada - por exemplo, a distância entre dois países, o nome de alguma pessoa, etc. As marcações atribuídas aos termos da pergunta oferecem restrições semânticas à eles, permitindo que estes sejam testadas em relação aos termos encontrados nos candidatos à resposta.

A eficácia desses sistemas, geralmente, está ligada ao nível de precisão de análise da pergunta. Nessa etapa são realizadas análises semântica e sintática, dentre outros métodos, podendo variar de acordo com a abordagem adotada no sistema, algumas delas são ontologias, WordNet, estatística, análise baseada em regras, entre outras (CARPINETO; ROMANO, 2012).

Em alguns casos, as palavras-chave extraídas da sentença são expandidas, usando alternâncias morfológicas, léxicas e semânticas. Esse processo permite que a extração de texto contenha conceitos chaves que não foram expressados explicitamente na pergunta (VICEDO; MOLLÁ, 2001).

A partir da identificação do foco e de outros elementos da pergunta, é gerada uma expressão contendo as palavras-chave que serão utilizadas num mecanismo de busca e recuperação de documentos. A etapa de seleção e extração da resposta utiliza as informações adquiridas pelo primeiro componente para realizar uma seleção inicial de documentos que contenham possíveis respostas. Em vista do volume que esses sistemas têm que gerenciar, além das limitações impostas por um tempo de resposta razoável, a ação de seleção é realizada, geralmente, por mecanismos de Recuperação de Informação (RI) (VICEDO; MOLLÁ, 2001). O resultado é a redução de toda a base de dados a uma pequena coleção com os arquivos de texto que contenham a resposta procurada.

A atividade de extração realiza uma análise detalhada dos textos relevantes selecionados, a representação da pergunta e dos textos relevantes são comparados entre si afim de obter o conjunto de respostas candidatas. O sistema então faz um ranking dos candidatos, baseado na probabilidade de exatidão da resposta e as apresenta ao usuário da forma mais adequada.

Os métodos de ranking são inúmeros, indo desde a frequência das palavras-chave, a distância entre termos presentes na pergunta e nos candidatos, entre outras. Geralmente esses métodos combinam técnicas de similaridade, popularidade (frequência), padrões específicos e validação da resposta (VICEDO; MOLLÁ, 2001).

A base de conhecimento trata do conjunto de dados de onde serão selecionados os documentos candidatos a resposta. A estrutura e o tipo de dados podem ir desde vastas coleções de texto, banco de dados com pares pergunta-resposta - no caso de sistemas do tipo FAQ - até arquivos de imagens, vídeos, entre outras. A tabela 3.2 apresenta, resumidamente, a relação entre os tipos de sistemas, o domínio e a base de conhecimento.

Tabela 3.2: Relação entre tipos de sistemas e bases de conhecimento, adaptado de Rodrigues (2007)

Tipos de sistemas	Domínio	Base de conhecimento
Interfaces para base de dados	Fechado	Estáticas
Baseado em enciclopédias	Fechado ou aberto	Geralmente estáticas
Baseado em coleções de texto não estruturadas	Fechado ou aberto	Estáticas (coleções locais de documentos) ou dinâmicas (Internet)

O componente de geração de resposta, geralmente, manipula textos. Contudo, os sistemas podem fazer uso de outros métodos de apresentação da resposta ao usuário, em formato de diálogo por exemplo. Em algumas abordagens, o sistema pode considerar o perfil do usuário para customizar a forma de apresentação, esse aspecto e demais características dos sistemas de perguntas e respostas são discutidos nas seções seguintes.

3.2 Características dos Sistemas QA

Nesta seção são apresentadas as características relacionadas à Sistemas de Perguntas e Respostas, tais como tipo de usuário, classificação dos tipos de perguntas, obtenção e construção da resposta, avaliação da resposta e formas de apresentação da resposta, além dos problemas inerentes à esta área.

3.2.1 Usuários

O tipo de usuários pode variar de acordo com diversos fatores, podendo ser nível de escolaridade, frequência de acessos, número de respostas corretas, entre outras. Com essa divergência, é necessário que o sistema atenda as diferentes necessidades de cada usuário.

BURGER et al. (2001) afirmam que o aprofundamento na compreensão do foco da questão varia em diferentes graus, de acordo com o nível de sofisticação dos usuários perguntadores, bem como as características do perfil de cada perguntador, como visto na figura 3.1. Na tabela 3.3 estão descritos os quatro níveis propostos.

Tabela 3.3: Características do perfil dos usuários perguntadores, adaptado de Burger et. al. (2001)

Nível	Características
1 – “Perguntador Casual”	Satisfação do usuário, frequência de perguntas, número de perguntas sequenciais, domínios de interesse, número de perguntas repetidas, número de perguntas relacionadas à um tópico
2 – “Perguntador Padrão”	Número de padrões relacionados, número de perguntas recuperadas e sua relevância, frequência de reutilização de padrões, frequência na introdução de novos padrões.
3 – “Repórter Novato”	Tamanho dos diálogos, tópicos de diálogos, complexidade do contexto, número de perguntas não respondidas, número de perguntas reformuladas, número de perguntas similares ou relacionadas numa mesma sessão.
4 – “Analista de Informação Profissional”	Número de novos fatos descobertos, relações, uso de feedback, tempo gasto navegando novas evidências, número de novos axiomas adicionados na base de conhecimento, requer esquemas complexos de compreensão

O perguntador casual realiza pergunta simples que exigem respostas diretas, geralmente encontradas em frases curtas em apenas um documento. Já o perguntador padrão tem modelos estabelecidos de perguntas, que necessita de contexto relevante na resposta procurada. O tipo de usuário com perfil de "repórter novato" procura por fatos, com o intuito de se aprofundar em um aspecto do tópico explorado, podendo necessitar de informações encontradas em diversos documentos. O analista de informação profissional procura um nível de contexto e aprofundamento elevado.

Usuários respondedores também necessitam de adaptações que variam de acordo com seu perfil. Para usuários iniciantes, pode ser importante explicar as limitações do sistema, para que estes saibam interpretar as respostas obtidas, já para usuários mais experientes pode ser útil a utilização de modelos, para focar em informações inéditas à eles (HIRSCHMAN; GAIZAUSKAS, 2001).

3.2.2 Perguntas

Em LEHNERT (1977) destacou a necessidade de categorização dos tipos de perguntas em um sistema de perguntas e respostas. Segundo a autora, quando deseja-se classificar todas as possíveis perguntas de um domínio, existem diversas maneiras de analisar a pergunta, por isso é importante categorizar a entrada de acordo com os procedimentos que serão acionados para respondê-la.

A autora propôs uma classificação baseada nas representações conceituais que estão subjacentes à língua inglesa, onde divide as perguntas em cinco (05) categorias: (1) *why questions* - perguntas que incorporam senso de causa; (2) *how questions* - conectam um ato ou ação determinada na pergunta à algum tipo de instrumentalização ou condição; (3) *yes or no questions* - procuram estabelecer a veracidade de uma afirmação; (4) *occurrence questions* - similar à segunda categoria, lida com a consequência seguinte à ação executada; e (5) *component questions* - ocorrem quando um componente atômico da conceptualização é desconhecido ou não identificado.

VICEDO; MOLLÁ (2001) dividiram o conjunto de possíveis perguntas baseados no tipo de resposta esperados:

- Perguntas factuais: requerem um ou mais itens específicos de informação, como datas, nomes de entidades, quantidade, etc. Por exemplo: *"Quanto é 2 + 2?"*
- Perguntas de síntese: necessitam que o sistema localize instâncias específicas de informação e as apresente de forma resumida, essas perguntas variam de listas até perguntas com estrutura pré-determinada. Por exemplo: *"Liste os filmes dirigidos por Quentin Tarantino"*
- Perguntas de contexto: estas perguntas são postadas como contexto de perguntas anteriores, dessa forma a interpretação da pergunta depende do significado e das respostas anteriores. Por exemplo: (1) *"Qual foi a primeira seleção a ganhar a copa do mundo?"* (2) *"Quem foi o adversário?"* (3) *"Qual foi o placar?"*
- Perguntas especulativas: são perguntas muito complexas que necessitam de coleções de dados e de técnicas dedutivas. Por exemplo: *"O que aconteceria com o Brasil se a reserva da Cantareira secasse completamente?"*

Perguntas factuais (factoides) ainda podem ser subdivididas em quatro categorias, que remetem à classificação proposta por LEHNERT (1977), com algumas atualizações propostas por VICEDO; MOLLÁ (2001): (a) perguntas do tipo sim ou não; (b) perguntas com pronomes interrogativos (perguntas wh- (*what, who, when*, etc); (c) perguntas indiretas; e (d) perguntas de comando - *"Me diga onde fica a cidade de Imperatriz"*.

MOLDOVAN et al. (1999) propuseram taxonomias para seu sistema proposto, como forma de determinar os tipos de pergunta e respostas em um sistema de domínio aberto. A figura 3.4 ilustra a taxonomia proposta.

Classe	Subclasse	Tipo de resposta esperada
what	basic what	DINHEIRO / NÚMERO / DEFINIÇÃO / TÍTULO NNP / INDEFINIDO
	what-who	PESSOA / ORGANIZAÇÃO
	what-when	DATA
	what-where	LOCALIZAÇÃO
who		PESSOA / ORGANIZAÇÃO
how	basic how	MANEIRA
	how-many	NÚMERO
	how-long	TEMPO / DISTÂNCIA
	how-much	DINHEIRO / PREÇO
	how-much-modifier	INDEFINIDO
	how-far	DISTÂNCIA
	how-tall	NÚMERO
	how-rich	INDEFINIDO
	how-large	NÚMERO
where		LOCALIZAÇÃO
when		DATA
which	which-who	PESSOA
	which-where	LOCALIZAÇÃO
	which-when	DATA
	which-what	NNP / ORGANIZAÇÃO
name	name-who	PESSOA / ORGANIZAÇÃO
	name-when	LOCALIZAÇÃO
	name-what	TÍTULO / NNP
why		RAZÃO
whom		PESSOA / ORGANIZAÇÃO

Figura 3.4: Taxonomia para perguntas proposta por Moldovan et. al. (1999)

3.2.3 Respostas

Para responder a uma pergunta, o sistema precisa analisá-la de acordo com o contexto. O sistema precisa encontrar uma ou mais respostas consultando fontes online e precisa apresentá-la ao usuário de forma apropriada (HIRSCHMAN; GAIZAUSKAS, 2001).

O tipo de resposta não segue um padrão, pode ser uma lista ou narrativa. Pode também variar de acordo com o tipo de usuário, caso este tenha interesse na justificativa ou contexto aprofundado.

Ainda de acordo com HIRSCHMAN; GAIZAUSKAS (2001), existem diferentes metodologias para a construção da resposta: por extração (cortando trechos de texto do documento original) ou geração. O método de geração consiste na integração de diferentes respostas retiradas de um ou mais documentos, nesse caso a coerência da resposta pode ser reduzida, o que consiste em um dos problemas pertinentes dos Sistemas de Perguntas e Respostas.

Nas conferências TREC e CLEF, em Sistemas de Perguntas e Respostas para perguntas factuais, cada candidato a resposta é avaliado como MAGNINI et al. (2006):

- Correto: se nenhuma informação é necessária além da apresentada. A resposta deve vir acompanhada do código de identificação do(s) documento(s) no qual(ais) a exata resposta foi encontrada; o documento precisa ser relevante;
- Não suportado: se o código de identificação não for encontrado ou estiver errado, ou o trecho extraído não contém a resposta exata;
- Inexato: se o trecho extraído necessita de menos ou mais informações apresentadas;
- Incorreto: se a resposta não provê a resposta necessária.

3.2.4 Avaliação

Para avaliar a qualidade de uma resposta faz-se necessário a seleção de critérios para avaliação, podendo ser avaliação automática ou através dos usuários. HIRSCHMAN; GAIZAUSKAS (2001) levantaram alguns possíveis critérios:

- Relevância: a resposta deve ser adequada à pergunta;

- Exatidão: a resposta deve ser factualmente correta;
- Concisão: a resposta não deve conter informação estranha ou irrelevante;
- Completude: a resposta deve ser completa, isto é, uma resposta parcial não deve ser considerada;
- Coerência: a resposta deve ser coerente, para que o usuário a leia facilmente;
- Justificativa: a resposta deve conter contexto suficiente para o leitor compreendê-la e o motivo dela ter sido escolhida.

Além de critérios de avaliação automática, as respostas podem ser avaliadas por leitores humanos através de testes de compreensão. Esse método, no entanto, requer que as respostas sejam avaliadas por mais de um leitor, por questões de consistência. Os critérios para seleção dos usuários avaliadores variam de acordo com a abordagem do sistema.

3.2.5 Principais Problemas

Alguns dos desafios inerentes à linguagem natural é a compreensão da ambiguidade e expressividade de algumas palavras, que podem ter diferentes significados em diferentes domínios (DAMLJANOVIC; AGATONOVIC; CUNNINGHAM, 2012). Por exemplo, a palavra "*área*" pode significar espaço, superfície ou campo de pesquisa, dependendo do contexto da pergunta e domínio do sistema. Uma das formas de controlar esse problema é a restrição da linguagem do sistema, como aplicado no sistema de LOPEZ et al. (2007).

O AquaLog, sistema proposto por LOPEZ et al. (2007), conta com um mecanismo de *plugin* que possibilita que o sistema seja configurado para diferentes representações de domínio. Essa abordagem permite que o sistema seja independente de domínio e, durante sua execução, atue apenas sobre uma representação linguística.

ZADEH (2006) acrescenta outros obstáculos para Sistemas de Perguntas e Respostas, todos relacionados à natureza humana como percepção, dedução e relevância. Percepção está, geralmente, relacionada ao conhecimento comum, que é adquirido através de experiência, comunicação e educação. Como, por exemplo,

"Janeiro é um dos meses mais quentes em Vitória", conhecimento que foi adquirido com a experiência e percepção do observador.

O problema da percepção em sistemas computacionais é que ela é intrinsicamente imprecisa, refletindo o fato de que o cérebro humano tem grande capacidade em resolver detalhes e reter informações.

Relevância exerce papel fundamental em qualquer busca. Conforme destacado por BRIN; PAGE (1998), o sucesso inicial do Google deve-se ao seu algoritmo de ranking de páginas baseado em sua relevância.

Com relação à parte técnica dos Sistemas QA, BURGER et al. (2001) levantaram outras dificuldades, apresentadas na tabela 3.4.

Tabela 3.4: Dificuldades técnicas em Sistemas QA, adaptado de Burger et. al. (2001)

Classes de Perguntas	Necessita de taxonomias para perguntas
Processamento da Pergunta	Compreensão, ambiguidade, implicações e reformulações de perguntas
Contexto	Definição, navegação e navegação de contexto
Fontes de Dados	Definir bases de dados disponíveis
Extração da Resposta	Extração de respostas simples e distribuídas; justificação da resposta e avaliação da exatidão
Formulação da Resposta	Apresentação de forma mais natural possível
Resposta em tempo real	Sistemas extraem respostas em tempo real, independente da escala da base de dados
Sistemas multilíngue	Sistemas independentes de língua
Interatividade	Modelo computacional de linguagem natural interativo, componente de diálogo baseado em análises extensas de amostras de diálogo
Raciocínio aprofundado	Integração de componentes de raciocínio, codificando conhecimento comum e mecanismos de conhecimento específico
Perfil do usuário	Modelagem do perfil do usuário

Recomendação	Identificação e sugestão de usuários ou perguntas sob o mesmo contexto

3.3 Trabalhos Correlatos

Nesta seção são apresentados os trabalhos analisados que serviram de base para o desenvolvimento do sistema descrito nesta dissertação. Para a seleção das publicações optou-se por trabalhos que aplicassem técnicas de Processamento de Linguagem Natural, tanto para a interpretação da pergunta quanto para a formulação da resposta, e que contam com uma porcentagem significativa de precisão de perguntas respondidas. O intuito da análise dos trabalhos aqui listados é aplicar algumas das técnicas utilizadas pelos autores no desenvolvimento tanto do AICAPA quanto do Agente de Dúvidas.

Sistemas como o *Yahoo! Answers*, *Quora* e *StackOverflow* apesar de lidarem, diretamente, com perguntas inseridas em linguagem natural, não aplicam técnicas de processamento para respondê-las. As perguntas são exclusivamente respondidas por outros usuários, sendo que não existem restrições sobre quem pode responder à alguma pergunta, existem apenas recursos de ranking, que destacam perguntas, categorias e/ou usuários mais populares ou mais frequentemente acessados.

Sob essa perspectiva, considerando as especificidades do problema e a abordagem adotada por cada trabalho proposto, as publicações selecionados são os que segue: ANDRADE et al. (2003), AMORIM; CURY; MENEZES (2011), DAMLJANOVIC; AGATONOVIC; CUNNINGHAM (2012) e GUO; ZHANG (2008).

3.3.1 QSabe

O trabalho de ANDRADE et al. (2003), MENEZES; TAVARES; PESSOA (1998) e PESSOA (1997) apresenta o QSabe, um ambiente virtual de perguntas e respostas, no qual cooperam e colaboram vários participantes com o objetivo de adquirir e compartilhar conhecimento. As interações com o ambiente permitem ao QSabe conhecer o tipo de usuário e, assim, aprimorar as atividades de escolha dos candidatos a respondedores. Os usuários podem desempenhar diferentes papéis no

sistema, podendo ser: respondedor, avaliador ou coordenador de contexto. A arquitetura desse sistema está ilustrada na figura 3.5.

No ambiente, as perguntas são encaminhadas aos usuários respondedores baseadas no cálculo de similaridade entre o perfil do respondedor e o contexto da pergunta, e reputação do usuário. A reputação é calculada usando a frequência de cada participante e o número de respostas bem avaliadas.

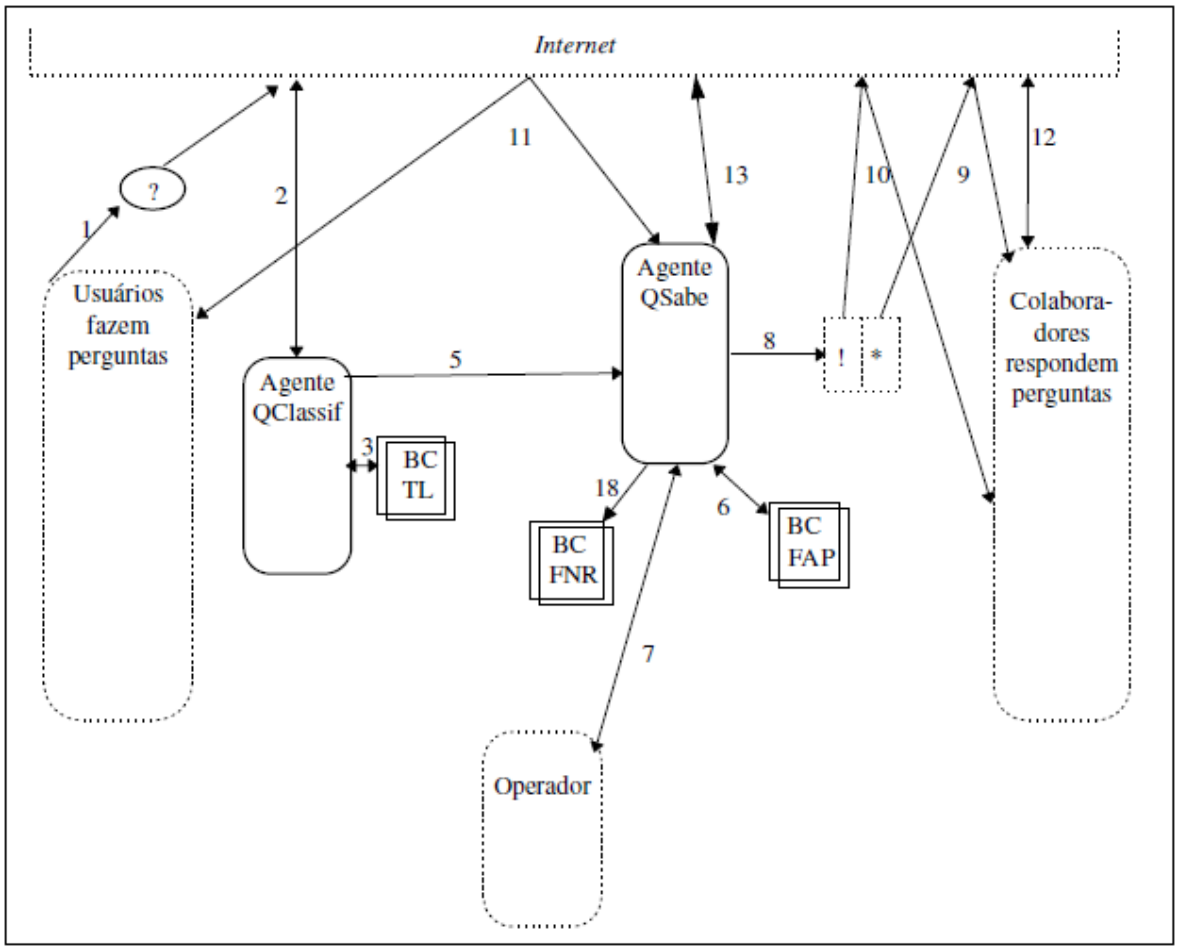

Figura 3.5: Arquitetura QSabe, retirado de (MENEZES; TAVARES; PESSOA, 1998)

Ao receber uma pergunta, o QSabe a armazena em um banco de perguntas, esta é então encaminhada aos candidatos a respondedores. O processo de classificação do tipo de pergunta é feito pelos usuários, ao enviá-la ao ambiente. Cada contexto (tipo de pergunta) é gerenciado por um usuário denominado coordenador, que pode reclassificar uma pergunta.

O usuário recebe várias respostas e pode utilizá-las da forma que melhor lhe convir, podendo escolher apenas uma ou somar diversas respostas diferentes. As

respostas passam por uma validação social, onde são avaliadas por um diferente participante especialista. As perguntas e suas respectivas respostas são apresentadas em um mural, onde todos os participantes podem visualizá-las, dessa forma, além de promover o conhecimento compartilhado, a possibilidade de outros usuários enviar perguntas já respondidas é reduzida.

3.3.2 Sistema de Esclarecimento de Dúvidas proposto por Amorim et. al. (2011)

Em AMORIM; CURY; MENEZES (2011), é proposto uma arquitetura de um sistema que visa receber perguntas e oferecer respostas de forma automática dentro do domínio de Sistemas Operacionais. O sistema é apoiado por ontologias, agentes de software e um banco de conhecimento AIML (*Artificial Intelligence Markup Language*). A arquitetura é dividida em dois grandes blocos: aquisição de conhecimento e responder perguntas, como ilustrado pela figura 3.6.

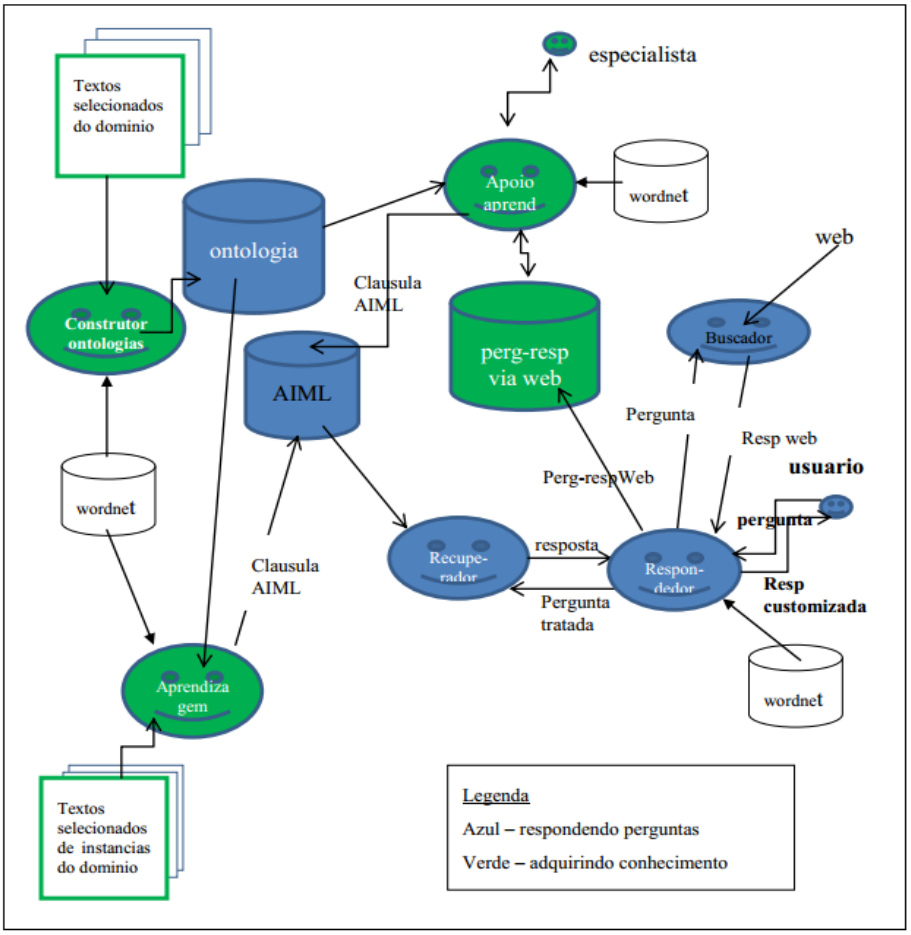

Figura 3.6: Arquitetura do sistema proposto por Amorim et. al., 2011

Nesse sistema, a aquisição de respostas ocorre de duas maneiras, ao receber uma pergunta esta é enviada ao agente respondedor, que a analisa através do WordNet, para questões de sinonímia. Então, a pergunta é enviada ao banco AIML, que busca por uma resposta adequada e, em caso positivo, a apresenta ao usuário. Caso a resposta não seja encontrada no banco AIML, o agente respondedor recebe a natureza semântica da pergunta e a envia ao agente buscador da web. Ao receber uma resposta da web, o agente respondedor envia a resposta ao usuário e a armazena no banco AIML, juntamente com a pergunta original.

O sistema adquire conhecimento de forma multinivelada, em um dos níveis é construída a ontologia de domínio, a partir de um conjunto de textos sobre o domínio. Em outro nível, as cláusulas do banco AIML são construídas, baseadas em assuntos específicos. Um terceiro nível trata da aquisição continuada de conhecimento.

3.3.3 Sistema QA proposto por Guo e Zhang (2008)

O trabalho publicado por GUO; ZHANG (2008), descreve um sistema de perguntas e respostas automático, baseado em ontologias e web semântica, dividido em três módulos: módulo de compreensão semântica da pergunta, módulo de similaridade da pergunta (baseado em FAQ) e módulo de armazenamento para recuperação automática da resposta.

Para identificação da pergunta inserida, o sistema utiliza uma combinação de técnicas de processamento de linguagem natural, tais como segmentação, marcação *part-of-speech* e extração de palavras-chave.

As respostas são selecionadas a partir de um modelo de similaridade baseado em FAQ, que é composto por uma análise de similaridade semântica entre os termos que compõe a pergunta e a possível resposta. Para a busca da resposta, o sistema lida inicialmente com uma base de dados textual, construindo um index reverso, em seguida faz uso de um modelo de recuperação de informação para procurar por documentos relevantes na base.

O módulo de extração da resposta conta com técnicas de recuperação de documentos, recuperação de trechos e correspondência de resposta. O sistema provê um método que calcula o peso dos termos e seleciona a resposta adequada.

3.3.4 FreYa (Damljanovic et. al., 2012)

Já o trabalho publicado por DAMLJANOVIC; AGATONOVIC; CUNNINGHAM (2012) propõe um sistema interativo que traduz uma pergunta em linguagem natural em uma consulta SPARQL, uma linguagem de consulta para RDF, o FreYa. Trata-se de uma interface de linguagem natural que consulta ontologias que combinando métodos de melhoramento de usabilidade, como feedback e diálogos de clarificação na tentativa de aperfeiçoar o recall e a precisão das respostas.

As opções de resposta apresentadas ao usuário são encontradas através de raciocínio de ontologia e são, inicialmente, ranqueadas usando uma combinação de similaridade de sentenças (*strings*) e detecção de sinônimos, feita com auxílio do WordNet. O sistema ainda conta com um método de aprendizado, que aprende as seleções do usuário, melhorando sua performance com o tempo.

3.3.5 Tabela Comparativa

Na tabela 3.5 estão dispostas, de forma resumida, as informações acerca das técnicas e métodos utilizados pelas publicações supracitadas, divididas nos principais módulos que compõe um Sistema de Esclarecimento de Dúvidas.

Tabela 3.5: Tabela comparativa - Sistemas QA

Sistema	Análise de Pergunta	Seleção, Extração e Geração da Resposta	Base de Conhecimento
QSabe Andrade et. al. (2003)	Manualmente (classificação feita pelo usuário coordenador)	Seleção dos perfis de usuário respondedor	Especialistas
Amorim et. al. (2001)	WordNet, ontologias, Banco AIML, expressões regulares	Seleção por grau de relevância, RTE	AIML / Web
Guo e Zhang (2008)	Segmentação, POSTagger, palavras-chave	Cálculo de similaridade baseado no peso dos termos, index reverso	Base textual
Damljanovic et. al. (2012)	Ontologias, WordNet	Consulta SPARQL, aprendizado de máquina, combinação de similaridade	Base textual

3.4 Considerações finais do Capítulo

Este capítulo discorreu sobre os Sistemas de Esclarecimento de Dúvidas, algumas das definições propostas para esses sistemas, classificações e características dos componentes básico da arquitetura.

Fizemos, também, uma análise das publicações relacionadas à área, que foram utilizadas como referência para a concepção e modelagem da proposta do Agente descrito nessa dissertação de mestrado.

O Agente de Dúvidas, em particular, contribui à área de QAS unindo métodos científicos, tais como algoritmos de Processamento de Linguagem Natural, e aspectos sociais em apoio à atividade de esclarecimento de dúvidas e compartilhamento de informação.

4 PROPOSTA DE SOLUÇÃO

O Agente de Dúvidas é o componente do AICAPA responsável por apresentar respostas adequadas aos questionamentos enviados pelos usuários. A partir da pergunta inserida, o Agente transforma a entrada em palavras-chave que compõe uma consulta a ser executada automaticamente sob uma base de dados, no intuito de apresentar a resposta mais adequada de acordo com o tipo da pergunta. A seção 4.1 apresenta a visão geral do Agente de Dúvidas, abordando esse objetivo geral, características principais e a arquitetura proposta para esse sistema.

Para realizar esta tarefa é necessário a execução de uma cadeia de processos, que envolvem Processamento de Linguagem Natural e Recuperação de Informação, o fluxo de atividades executadas é apresentado na seção 4.2.

4.1 Visão Geral

O Agente de Dúvidas corresponde à um Sistema de Esclarecimento de Dúvidas, desenvolvido com o objetivo de apresentar, de forma automática, respostas adequadas aos questionamentos inseridos, em linguagem natural, pelos usuários. A figura 4.1 apresenta a modelagem conceitual do funcionamento do Agente.

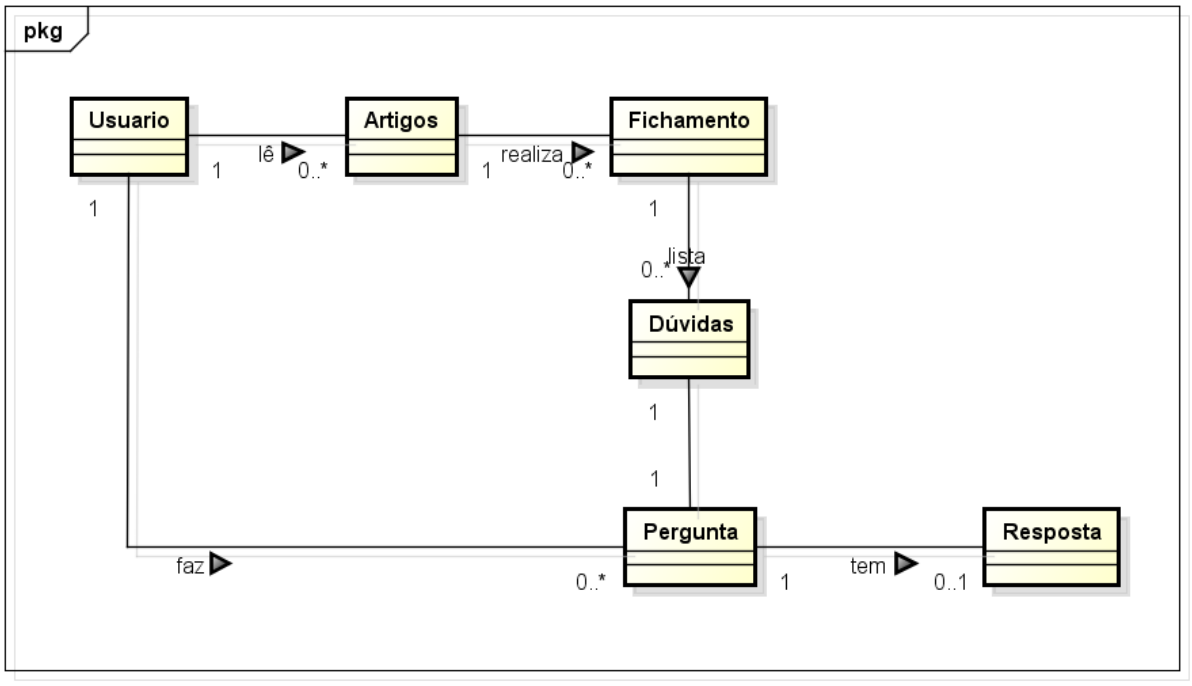

Figura 4.1: Modelagem Conceitual

Como ilustrado na figura 4.1, o Agente de Dúvidas recebe questionamentos de duas maneiras: através de envio de perguntas diretamente do usuário e identificando as dúvidas listadas nos fichamentos realizados pelo usuário. Ao realizar a leitura de um artigo, o usuário pode elaborar fichamentos no AICAPA, um dos modelos disponibilizados pelo ambiente conta com um campo para dúvidas surgidas ao longo da leitura. Essas dúvidas são identificadas pelo Agente, que as envia aos especialistas para responde-las.

O usuário também tem a possibilidade de enviar perguntas diretamente ao Agente, que então busca automaticamente por uma resposta adequada ao questionamento.

A proposta desse sistema é responder perguntas do tipo WH (o que, quando, quem, qual, quais) e listas simples dentro de um contexto determinado pelo usuário. Por exemplo, sob um contexto de Redes de Computadores, é possível fazer perguntas como: “O que é o modelo OSI?” ou ainda “Liste as camadas do modelo OSI”.

Para atender esse objetivo são necessárias diversas etapas de processamento, por esse motivo este componente foi dividido em quatro (04) módulos: módulo *query*, base de conhecimento, módulo de busca e módulo de apresentação, cada um com atividades e objetivos bem definidos. A nomenclatura dos módulos foi determinada seguindo a atividade desenvolvida por cada um. A arquitetura proposta e as interações entre os módulos estão representada na figura 4.2.

4.1.1 Módulo Query

O módulo *query* é responsável por analisar a pergunta, inserida em linguagem natural pelo usuário, extrair palavras-chave e construir a query de busca, como a nomenclatura do módulo sugere. Também encarregado da identificação e classificação do tipo de pergunta, o resultado da análise é determinante para a eficiência do sistema por completo, já que os possíveis candidatos à resposta serão baseados nas palavras e no tipo de resposta esperada.

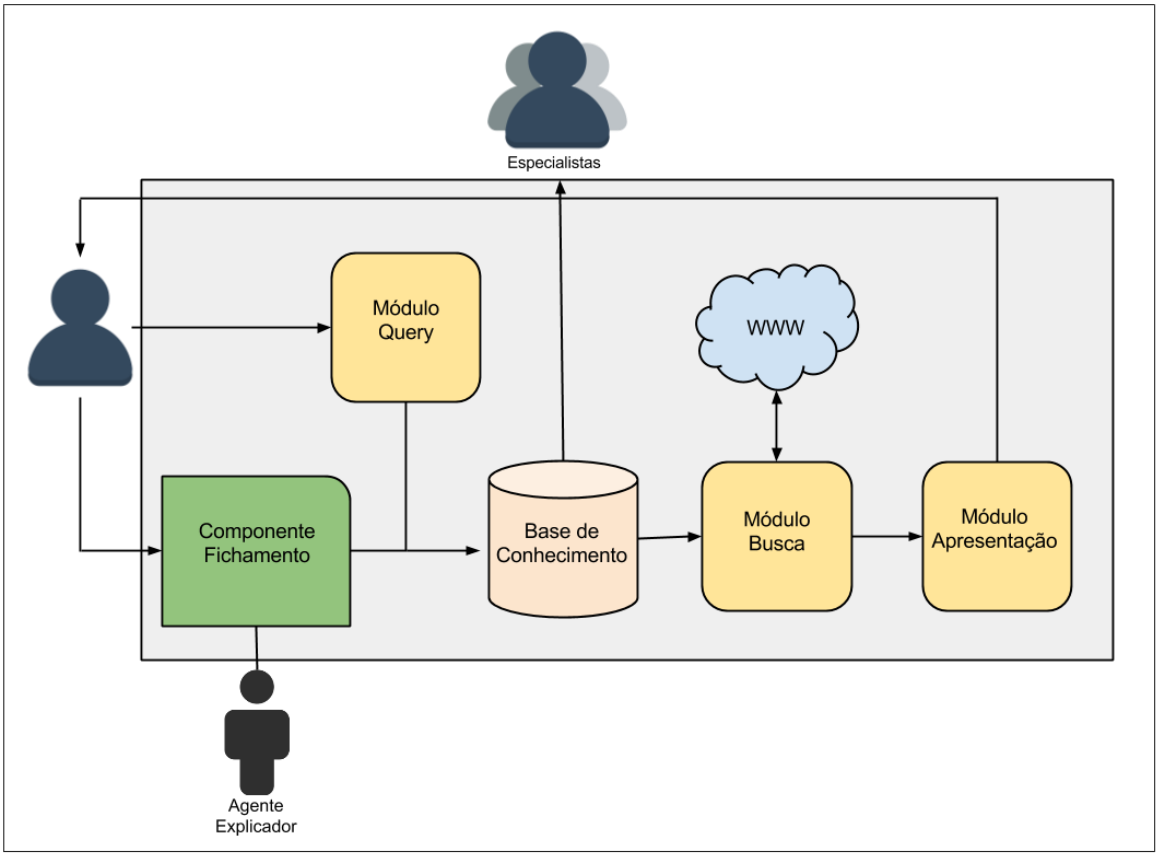

Figura 4.2: Arquitetura proposta

Para identificar o tipo semântico da pergunta, a entrada passa por etapas de pré-processamento, que compreendem as atividades de quebra da sentença em *tokens*, remover as palavras sem peso semântico (*stopwords*) e marcar cada *token* de acordo com a sua classificação gramatical.

Algumas perguntas, no entanto, contêm palavras que, apesar de não ser relevantes à determinação do tipo da pergunta ou também não irão compor a query de busca, servirão de parâmetros para a determinar o tipo de resposta esperada. Perguntas do tipo “o que, qual, quais” procuram por respostas relacionadas a definição, a entidades ou local. Já perguntas do tipo “quando” esperam respostas associadas a noções de tempo. Perguntas do tipo “quem”, têm candidatos ligados a pessoas. Listas representam um conjunto de informações, que pode ser retirado de um ou mais documentos. A partir da identificação e marcação dos *tokens*, o Agente seleciona o conjunto de palavras-chave que vão compor a query de consulta.

Em casos de perguntas não identificadas, devido à complexidade ou ausência de palavras de semântica relevante, estas são encaminhadas aos especialistas

humanos, determinados pela reputação no ambiente, para que estes possam oferecer respostas aos questionamentos. A reputação do usuário é determinada a partir das interações deste com o Ambiente, o número de respostas corretas, número de perguntas enviadas.

O caso de uso ilustrado na figura 4.3 apresenta as atividades desempenhadas pelo Módulo Query. Nesse diagrama, o módulo query foi representado como subsistema do Agente de Dúvidas.

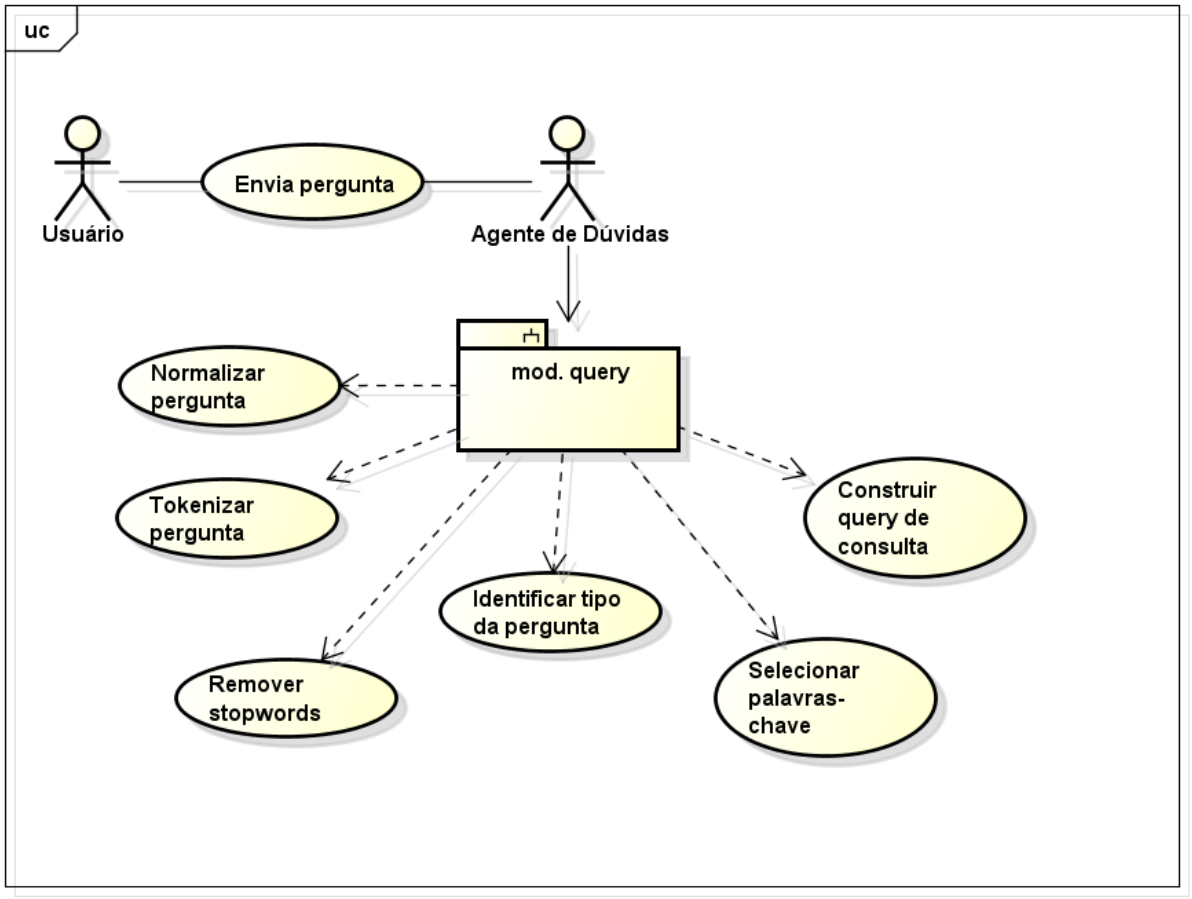

Figura 4.3: Caso de uso - Módulo Query

4.1.2 Base de Conhecimento

Este módulo da arquitetura é responsável pelo armazenamento dos pares pergunta-resposta enviados pelos módulos de query e de busca. Esses pares serão utilizados pelo sistema como base textual para busca de respostas para perguntas similares às questões solucionadas.

A base de conhecimento também armazena as dúvidas listadas pelo usuário no componente de fichamento. Estas dúvidas serão então encaminhadas ao Agente Explicador que encaminhará as dúvidas aos usuários especialistas.

Este banco de dados é consultado apenas pelo Agente de Dúvidas, ao contrário do que acontece com o Banco de Dados de Artigos, que é comum à todos os componentes do AICAPA.

4.1.3 Módulo de Busca

As atividades relacionadas à procura, recuperação e extração da resposta cabem ao módulo de busca. Uma das formas de busca realizadas por esse componente é consultar a base de conhecimento, onde estão armazenados os pares pergunta-resposta que foram respondidos em outras sessões. As palavras-chave retiradas formulam a query de consulta que será executada sob a base de conhecimento.

O outro método de busca e recuperação realizado por esse módulo é a consulta externa, fazendo uso de mecanismos de busca na Internet. A query formulada pelo módulo query é então executada, automaticamente, no módulo de busca. O resultado da consulta é a página de resultados recuperada pelo módulo, esta página é então analisada com o *parser*, no intuito de remover todas informações acerca de formatação da página que não estejam relacionadas à resposta procurada.

Com a remoção das formatações da página, o texto é analisado novamente pelo *parser*, afim de localizar os trechos de texto sobre as publicações recuperadas através da consulta.

A saída deste módulo são os trechos de texto que contém a resposta ao questionamento inserido, que são encaminhadas ao módulo de apresentação para exibi-las ao usuário de acordo com o tipo da pergunta.

Ao encaminhar a resposta ao Módulo de Apresentação, o módulo de busca envia o par pergunta-resposta para o banco de conhecimento e ao Agente de Interface, do AICAPA, que atualiza o perfil do usuário com o par correspondente.

A figura 4.4 apresenta as atividades que o Módulo de Busca executa para atingir o objetivo de buscar os candidatos a resposta. Assim como no caso de uso do

Módulo Query, o Módulo de Busca também foi representado como subsistema do Agente.

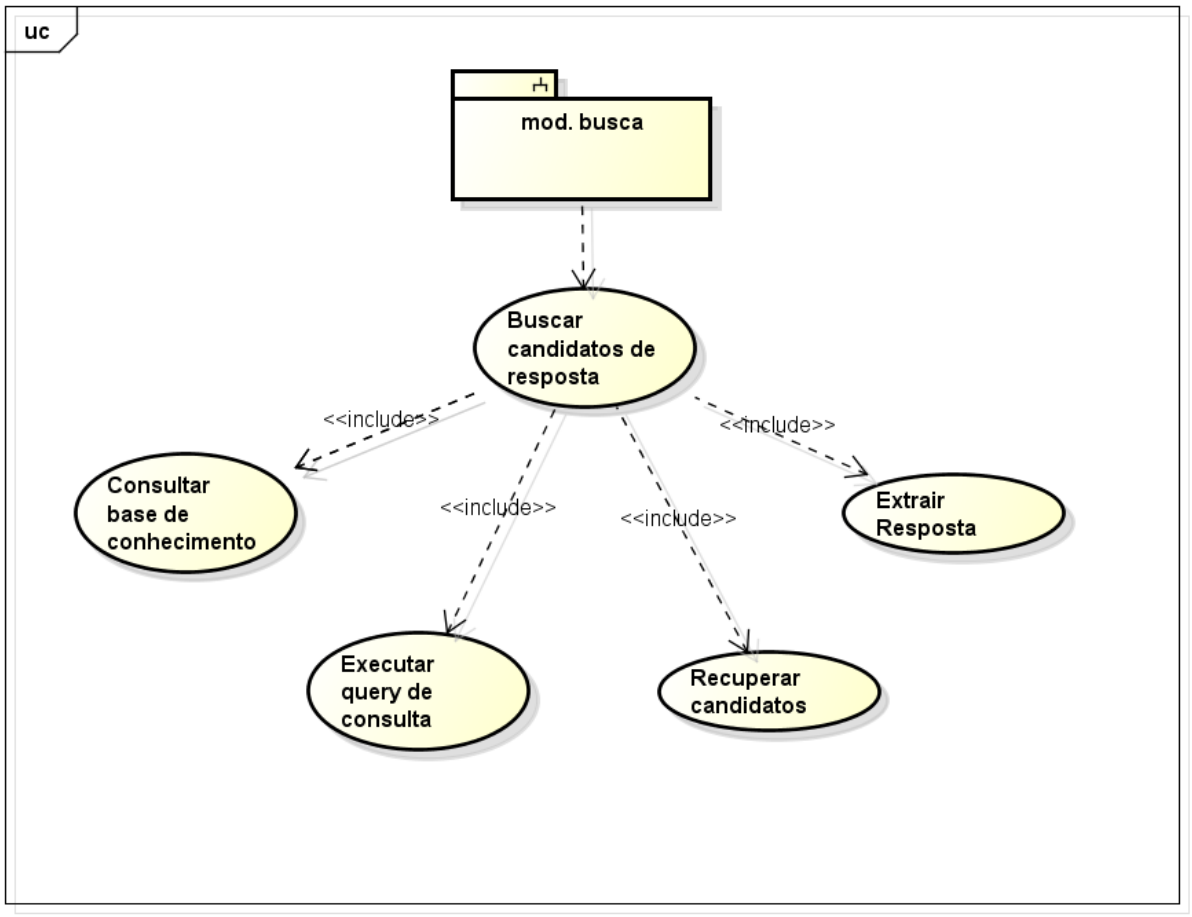

Figura 4.4: Caso de uso - Módulo de Busca

4.1.4 Componente de Fichamento

O componente de fichamento, apesar de estar representado na figura que ilustra a arquitetura do Agente de Dúvidas, não é módulo integrante deste Agente. A sua inclusão na arquitetura deve-se à interação do Agente de Dúvidas e do Agente Explicador com o campo de dúvidas presente em um dos modelos de fichamento disponíveis no AICAPA.

Como mencionado anteriormente, o campo de dúvidas pode ser preenchido pelo usuário no decorrer da leitura dos artigos selecionados. Ao enviar uma dúvida para o sistema, o Agente Explicador é acionado, que a identifica e envia o questionamento aos usuários especialistas.

No momento em que a pergunta for resolvida por um dos especialistas, o Agente Explicador envia o par pergunta-resposta para o banco de conhecimento do Agente de Dúvidas.

A figura 4.5 ilustra a relação entre o componente de fichamento, as perguntas e as respectivas respostas. De acordo com a modelagem, um fichamento pode não ter dúvidas ou pode ter mais de uma. Cada pergunta tem uma resposta adequada.

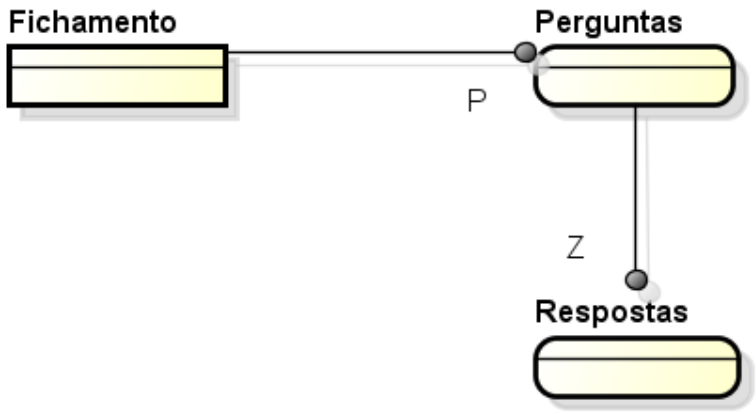

Figura 4.5: Diagrama Entidade Relacionamento - Fichamento

4.1.5 Módulo de Apresentação

O quarto módulo, apresentação, é encarregado de exibir a resposta obtida ao usuário, de acordo com o tipo de pergunta. A composição dos trechos exibidos é determinada, diretamente, pelo tipo de pergunta e a resposta esperada por ele. Nos casos de pergunta WH, são exibidas as informações procuradas de forma direta (pessoa, entidade, número) juntamente com o link da publicação recuperada – caso esteja disponível. Já com as listas apresentam todos os parâmetros de informação das publicações recuperados pelo *parser*.

O usuário tem a opção de fornecer informações acerca da qualidade da resposta, através de votação de adequação à pergunta inserida. As opções de votação são as classificações das respostas apresentadas no capítulo 3.

Esse módulo também aciona o Agente de Recomendação, do AICAPA, que exibirá um ou mais artigos utilizados para a composição da resposta, além de demais artigos relacionados à pesquisa do usuário.

4.1.6 Agente Explicador

Este Agente Explicador atua diretamente com o componente de Fichamentos do AICAPA. É acionado sempre que um usuário cadastra uma dúvida no modelo de fichamento do ambiente.

O agente, então, encaminha a dúvida para os usuários especialistas do ambiente. Quando o questionamento for resolvido, o Agente é acionado para enviar o par pergunta-resposta para o banco de conhecimento do Agente de Dúvidas e então atualizar o perfil do usuário. O caso de uso representado na figura 4.6 demonstra essa ação.

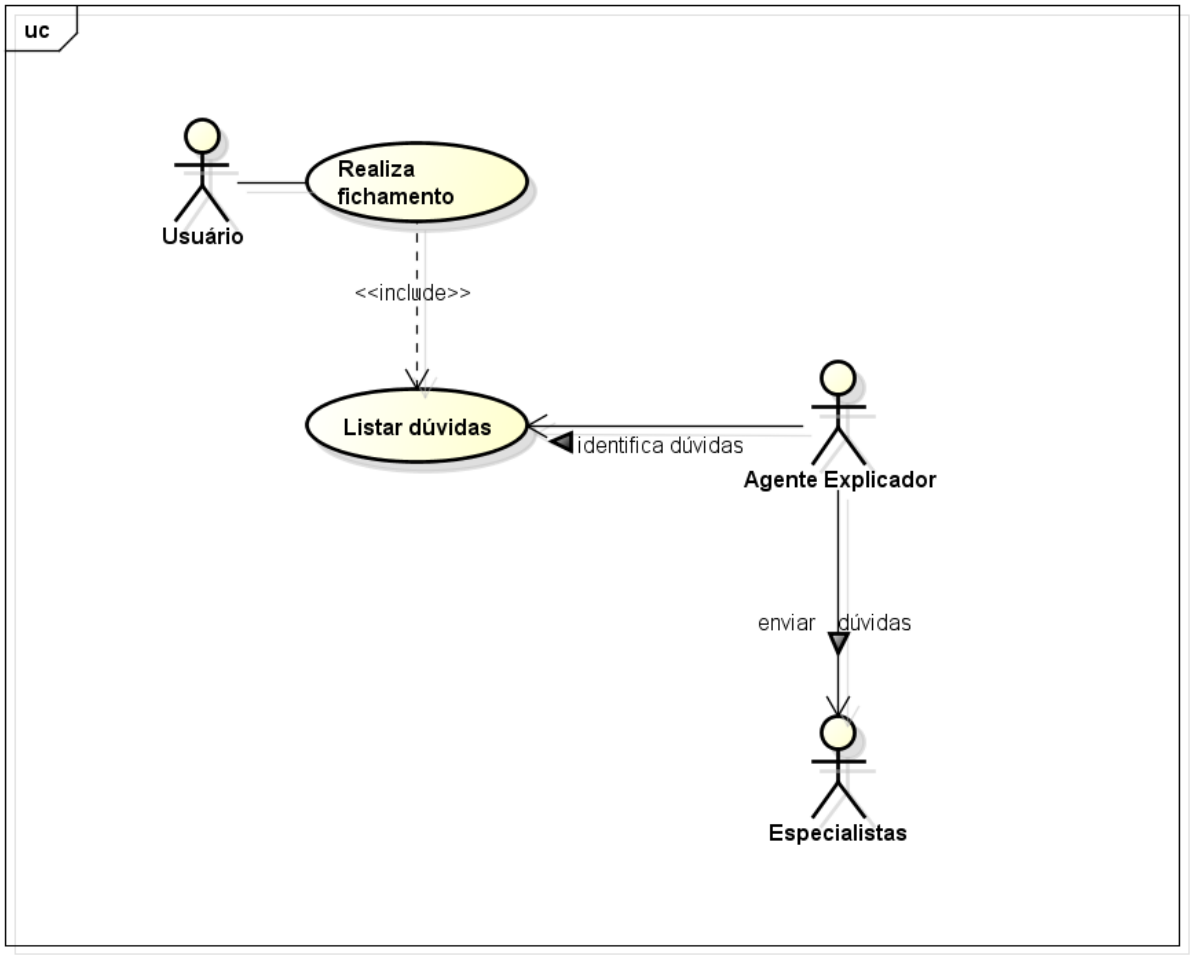

Figura 4.6: Caso de uso - Agente Explicador

4.2 Fluxo de Atividades

As atividades e interações entre módulos do Agente de Dúvidas apresentado neste trabalho são descritos nesta seção. São elas:

1. O Agente de Interface, do ambiente colaborativo AICAPA, recebe a pergunta em linguagem natural e a envia ao Agente de Dúvidas;
2. No Agente de Dúvidas, a pergunta é então analisada pelo módulo query, com o intuito de identificar o tipo semântico dos termos e o foco da pergunta;
3. Na atividade de análise, as palavras sem valor semântico (*stopwords*) são removidas, as palavras que tenham relação direta com a pergunta e a futura resposta, são consideradas como palavras-chave;
4. A query é então comparada com as palavras-chave dos pares (pergunta-resposta) armazenadas no banco de conhecimento, caso algum par seja compatível é apresentado ao usuário;
5. Em caso contrário, as palavras-chave são então transformadas numa query que será encaminhada ao módulo de busca, que executa a solicitação de busca.
6. O resultado da busca é então analisado pelo *parser*, no intuito de retirar toda a informação que não seja relevante à busca, ou seja, retirar todos os elementos de marcação da página (*tags* da formatação da página), entre outros.
7. De posse do excerto que representa a resposta, este é encaminhado ao módulo de apresentação, que então dispõe os elementos de texto extraídos de forma a facilitar a sua compreensão.
8. A resposta é encaminhada ao Agente de Interface, que aciona o Agente de Recomendação para apresentar recomendações de artigo e/ou grupos de discussão relacionados à pergunta respondida;
9. Em caso de ausência de resposta (o usuário votou como a resposta apresentada como inadequada), a pergunta é encaminhada aos usuários especialistas;
10. O perfil do usuário é, então, atualizado com o par pergunta-resposta.

4.3 Componentes Tecnológicos

Para o desenvolvimento do protótipo apresentado, optou-se por ferramentas livres, tanto pela gratuidade de acesso e utilização, como pela possibilidade de customização de códigos fonte e a disponibilidade de API (*Application Programming Interface*).

Os componentes tecnológicos aqui listados foram escolhidos baseados na extensão das bibliotecas disponíveis e a sua maturidade. Além da possibilidade de explorar uma tecnologia que não foi utilizada nos trabalhos selecionados para análise.

Nesta seção são descritas as funcionalidades de cada componente bem como o módulo onde foi utilizada.

4.3.1 Python

A linguagem de programação Python, criada por Guido Van Rossum em 1989, é um exemplo de linguagem de alto nível assim como C, Perl e Java. Em BIRD; KLEIN; LOPER (2009), Python é descrita como uma linguagem simples porém eficente com excelentes funcionalidades para o processamento de dados linguísticos.

Segundo ROSSUM (2008), Python é uma linguagem de alto nível, interpretada, orientada a objeto com semântica dinâmica. Seu alto nível construído em estrutura de dados, combinados com tipagem dinâmica, a tornam muito atraente para o desenvolvimento rápido de aplicativos, bem como para a utilização como uma linguagem de *script* ou para conectar componentes existentes.

Python oferece suporte à bibliotecas e pacotes que incentivam a modularidade do programa e reutilização de código. Além de suportar uma ampla gama de aplicações, desde simples *scripts* de processamento de texto até navegadores web interativos.

Atualmente, Python encontra-se na versão 3.4, contudo para o desenvolvimento da aplicação proposta nesse trabalho optou-se por usar a distribuição 2.7, por questões de compatibilidade com outras bibliotecas utilizadas na implementação do protótipo.

4.3.2 NLTK

Natural Language Toolkit (NLTK) é uma coleção de módulos e corpora registrada sob a licença *open source*, que permite que estudantes aprendam e conduzam pesquisas em processamento de linguagem natural (NLTK, 2013). O kit NLTK foi desenvolvido, inicialmente, em 2001 como parte do curso de Linguísticas Computacionais do Departamento de Ciência da Computação e Informação na Universidade da Pensilvânia (BIRD; KLEIN; LOPER, 2009).

É uma plataforma para construir programas Python que trabalham com dados em linguagem humana. Provê interfaces de fácil utilização para mais de 50 corpora e fontes léxicas como o WordNet, além de contar com uma suíte de bibliotecas de processamento para classificação, tokenização, *stemming,* POStagging, análise e raciocínio semântico.

Cada módulo do NLTK define uma estrutura ou tarefa específica (BIRD, 2006). Um conjunto de módulos centrais define tipos básicos de dados e sistemas de processamento que são utilizados por todo o kit. O NLTK conta com uma estrutura apropriada para o desenvolvimento de sistemas de processamento de texto em linguagem natural, alguns dos seus componentes principais são (RUSSELL, 2003):

- Definições de estrutura de dados que permitem que o computador represente entidades linguisticamente: *tokens*, árvores, regras gramaticais e distribuições de frequência;
- Classes e funções pré-programadas para diversas atividades de processamento de linguagem natural, incluindo: classificadores básicos de *n-grams*, *parsers* recursivos, *parsers* gráficos, e vários *parsers* probabilísticos.
- Uma série de representações gráficas. É possível extrair imagens estáticas, como árvores sintáticas ou histogramas básicos de distribuições de frequência.

Dados de amostragem (a sua maioria em inglês) para exercícios de prática, incluindo uma parte do *Penn Treebank*[6], entre outros.

4.3.3 Django

Django é um *framework* para aplicações web escrito em Python. Um framework Web provê uma infraestrutura simples para aplicações desse gênero, permitindo que o desenvolvedor mantenha o foco na criação de um código claro, manutenível e escalável. Django tem ganhado destaque, em comparação a outros frameworks Python (Zope, TurboGears) por proporcionar um meio simples, fácil e ágil para o desenvolvimento de sistemas para a internet. Django une a maturidade de Python e a portabilidade de sistemas web em um design simples e robusto.

[6] The Penn Treebank Project fornece estruturas de *parsers* que mostram informações sintáticas e semânticas (http://www.cis.upenn.edu/~treebank/)

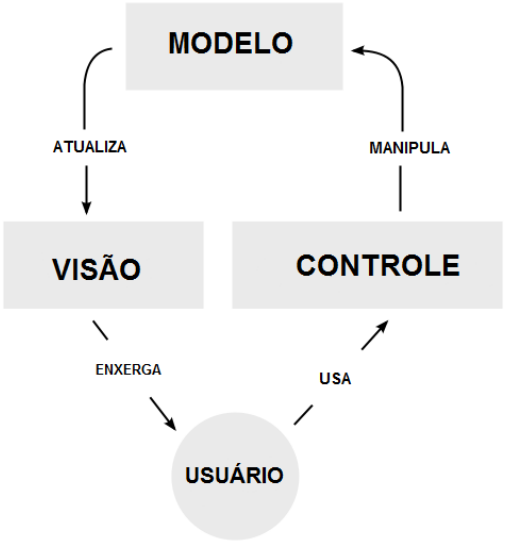

Figura 4.7: Padrão MVC

O desenvolvimento de aplicações com este framework segue o padrão, MVC (*Model-View-Controller*), que divide o projeto em três módulos: modelo, visão e controle, ilustrado na figura 4.7. No primeiro módulo, modelo, estão as definições dos dados, como eles serão armazenados e tratados, nessa camada representa quais os campos da tabela do banco de dados, os tipos de dados e seus valores padrão.

O componente de visão refere-se às funções que recebem requisições e retornam respostas ao usuário, a outro computador ou qualquer componente externo. Essa camada solicita informações ao modelo, que são então processadas e retornam alguma apresentação ao usuário, ou seja, representam a interface da aplicação. Já o módulo de controle recebe as entradas do sistema e as converte em comandos para o modelo ou visão.

Outro atributo deste framework é a forma como os sistemas são construídos, sua mobilidade e flexibilidade. No Django, os sistemas são criados como projetos, que podem ter inúmeros aplicativos em um mesmo projeto. De acordo com (HOLOVATY; KAPLAN-MOSS, 2009), um aplicativo no Django é uma aplicação Web que realiza uma tarefa, por exemplo um sistema de blog. Um projeto é uma coleção de configurações e aplicativos para um sistema Web em particular. Um projeto pode conter múltiplos aplicativos e um aplicativo pode estar em diversos projetos. Pode-se destacar também as interfaces de gerenciamento disponibilizadas por padrão, que

oferecem maior facilidade na manutenção dos dados manipulados pela aplicação Django.

O Django é uma ferramenta *open source*, distribuída sob a licença BSD, e pode ser adquirido gratuitamente no site *DjangoProject* sem qualquer restrição quanto ao sistema operacional.

4.3.4 Parser Google Scholar

Google Scholar (Google Acadêmico) é um serviço voltado para a busca e recuperação de publicações relacionadas à academia. Mesmo se tratando de uma ferramenta de livre acesso, as buscas diretas não podem ser incorporadas dentro de ambientes de terceiros, a desenvolvedora ainda não disponibilizou uma API oficial que permite a sua utilização como um serviço web.

Para fazer uso das informações desta ferramenta, Christian Kreibich desenvolveu, em Python, um *parser* das saídas de busca do Google Scholar (KREIBICH, 2013). O *parser* possibilita a extração de informações dos artigos listados nos resultados, como: título, URL, número de citações, entre outras, através de parâmetros de entrada semelhantes aos termos inseridos nas buscas na ferramenta online.

Contudo, o Google Acadêmico apresenta restrições quando utilizado através de aplicações de terceiros. Uma delas é o número de acessos, que são bloqueados quando detectado comportamento anômalo. O número exato de acessos não pôde ser determinado, varia de acordo com os parâmetros de pesquisa.

4.3.5 Beautiful Soup

O processo de varredura de páginas da Web é largamente utilizado como forma de obtenção de dados de web sites. Independente da informação, essas técnicas de varredura permitem a coleção de um grande volume de dados com o mínimo possível de esforço, além de dispensar o uso de base de dados ou outras formas de acesso às informações dispostas nas páginas web.

Beautiful Soup é uma biblioteca Python que analisa documentos nos formatos HTML e XML. O pacote percorre o documento, criando uma árvore dos dados analisados, podendo ser utilizados para a extração de informação desses documentos.

Utilizado de forma individual, o *Beautiful Soup* necessita que a página analisada seja armazenada localmente, para que então seja analisada. Em conjunto com outras bibliotecas, como o *urllib*, esse pacote pode buscar e extrair informações direto das páginas Web, sem a necessidade de armazenar as informações localmente.

4.3.6 Atividades da proposta com os componentes tecnológicos

A tabela 4.1 apresente, de forma resumida, as atividades desempenhadas pelos módulos do Agente e quais os componentes tecnológicos responsáveis.

Tabela 4.1: Atividades do Agente com os componentes tecnológicos

Atividade	Descrição	Componente tecnológico
Cadastros gerais	Inserir informações no ambiente de forma geral (cadastro de usuário, envio de pergunta, interações com o ambiente)	Django
Normalização	Transformar toda a entrada numa string de caracteres minúsculos, sem pontuação	NLTK
Tokenização	Dividir a sentença (*string*) que corresponde à pergunta em palavras.	NLTK
Remoção de *stopwords*	Remover as palavras que não tem peso semântico na construção da query.	NLTK
Palavras-chave	Selecionar as palavras que vão compor a query de busca	Python, NLTK
Construir a query	Formular a entrada da consulta	Python
Buscar candidatos a resposta	Busca por candidatos a resposta na base de conhecimento local	Python
Buscar candidatos a resposta	Busca por candidatos a resposta através da execução da query construída	Python

Análise do resultado	Varrer as páginas recuperadas através da query	Parser, Beautiful Soup
Seleção da resposta	Remover as informações que não fazem parte da resposta esperada	Parser, Beautiful Soup
Apresentação da resposta	Exibir o resultado da busca de acordo com a informação procurada	Parser, Python
Encaminhar perguntas aos especialistas	Encaminhar respostas que foram classificadas como inadequadas aos especialistas humanos	Python

4.4 Considerações finais do capítulo

Este capítulo apresentou os módulos da arquitetura da solução proposta e o funcionamento conceitual. O objetivo da arquitetura proposta é responder perguntas WH (o que, quando, quem, qual, quais) e listas simples dentro de um contexto determinado pelo usuário. A arquitetura foi dividida em quatro componentes bem definidos e, para atingir o objetivo, propomos a utilização de técnicas de Recuperação de Informação atuando sobre os mecanismos de busca.

5 PROVA DE CONCEITO

Neste capítulo são descritos os métodos, as técnicas e de que forma os componentes tecnológicos foram utilizados na implementação do recorte a proposta apresentada nessa dissertação.

A seção 5.1 descreve a implementação da proposta apresentada que consolida os conceitos e métodos expostos previamente nessa dissertação, abordando as técnicas aplicadas, como os componentes tecnológicos interagem entre si, no Agente de Dúvidas e no ambiente AICAPA.

A seção 5.2 apresenta os componentes tecnológicos utilizados na implementação do Agente descrito nesta dissertação.

Já a seção 5.3 apresenta os testes realizados com o protótipo desenvolvido e discorre sobre os resultados obtidos.

5.1 Protótipo

No intuito de testar a viabilidade da abordagem, foi selecionado um recorte da solução proposta como prova de conceito. A porção implementada neste protótipo corresponde às atividades de construção da query (análise da pergunta), busca, recuperação e extração da resposta (construção da resposta) e apresentação da resposta.

O protótipo do Agente de Dúvidas foi implementado na linguagem Python, com auxílio dos componentes tecnológicos listados na seção 5.2. Todo o Agente, conforme explanado no capítulo 3, corresponde à um Sistema de Esclarecimento de Dúvidas, e foi dividido em três (03) módulos: *query*, *search* e *presentation*.

A figura 5.1 apresenta as principais funções utilizadas na execução das atividades de cada componente do recorte implementado.

Foram elaboradas perguntas que podem servir de ponto de partida para a pesquisa do acadêmico, as perguntas utilizadas estão listadas no apêndice A. Na implementação dessa prova de conceito não serão contempladas as dúvidas listadas no componente de fichamento do AICAPA.

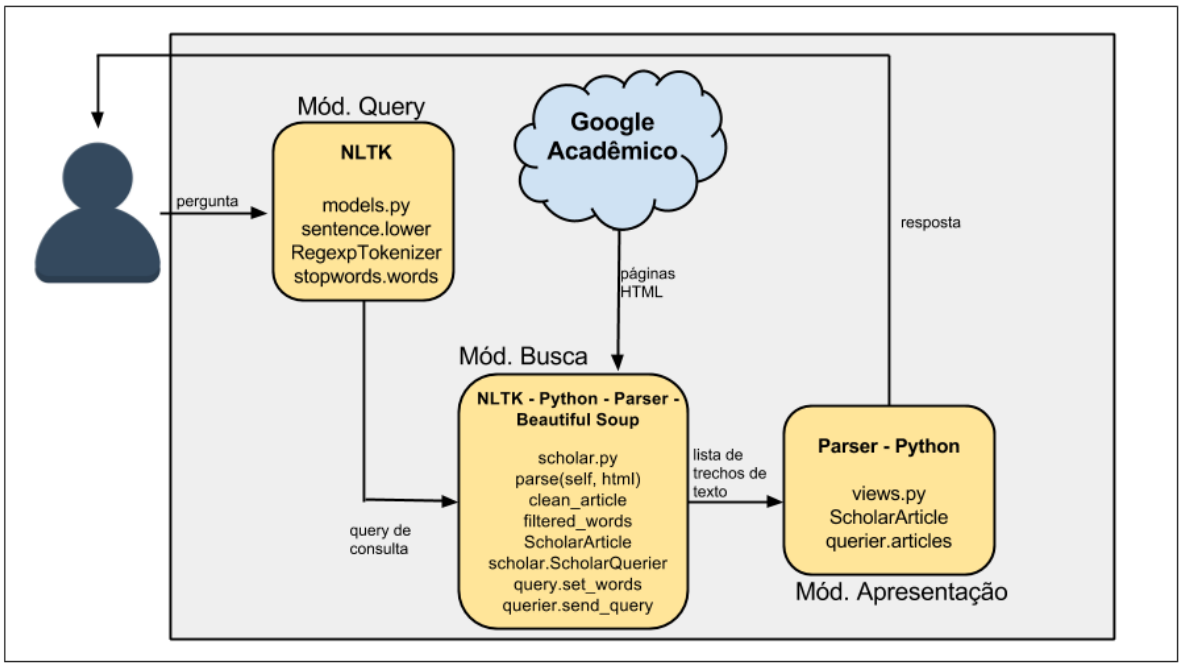

Figura 5.1: Arquitetura do recorte implementado

O protótipo foi desenvolvido para trabalhar com perguntas do tipo WH (*what, who, when*), e listas simples, ambas na língua inglesa. Estas listas podem ser consideradas simples por não serem construídas com informações extraídas de mais de uma fonte textual.

Optamos por utilizar a língua inglesa por questões de compatibilidade com as bibliotecas de análise linguística disponíveis nos componentes utilizados. A fonte de dados utilizada para o protótipo é o Google Acadêmico. Sob essa perspectiva, perguntas de caráter descritivo, tais como “o que é um sistema de esclarecimento de dúvidas?” não serão tratadas nessa implementação.

O protótipo inicia o funcionamento a partir do envio da pergunta, que é recebida pelo módulo query onde passa por uma série de processos de pré-processamento, com o intuito de remover as informações que não sejam relevantes à construção da query de consulta. O próximo passo é a execução desta query pelo módulo de busca, o resultado da consulta é uma página HTML contendo uma lista de publicações relacionadas à query.

O módulo de busca analisa a página, com um parser aplicado para o Google Acadêmico, removendo informações de formatação da página. Em seguida, o resultado dessa varredura é novamente analisado pelo parser, buscando o melhor

candidato para a consulta executada. O trecho escolhido é extraído e enviado ao Módulo de Apresentação, que apresenta o resultado de acordo com o tipo de resposta esperada.

A apresentação dessa implementação foi dividida em duas partes, interface e processamento, que são descritas nas seções seguintes.

5.1.1 Interface

Para o desenvolvimento da interface e seus componentes de interação foi utilizado o framework Django. Esse framework disponibiliza, de forma padrão, componentes para cadastro e gerenciamento de usuários, podendo ser customizado de acordo com as características da aplicação. A arquitetura do Django permite a acoplagem de outros aplicativos num mesmo ambiente, ou seja, em um mesmo sistema é possível integrar componentes independentes, que podem interagir entre si compartilhando informações.

No momento do cadastro, o usuário envia uma cópia do seu currículo Lattes no formato XML. A partir disso, o sistema localiza e extrai informações relacionadas a escolaridade, publicações e áreas de interesse do usuário. As palavras-chave destacadas nas publicações serão listadas como áreas de interesse do usuário, caso esse campo não esteja preenchido no arquivo XML, o usuário poderá acrescentar de forma manual.

Os campos relacionados as áreas de conhecimento também serão listadas como áreas de interesse. O recorte dessas informações no arquivo XML é apresentado na figura 5.2.

```
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Ambientes Colaborativos" PALAVRA-CHAVE-2="Revisão da Literatura" PALAVRA-CHAVE-3="Técnicas de Recomendação" PALAVRA-CHAVE-4="Técnicas de
Recuperação da Informação" PALAVRA-CHAVE-5="" PALAVRA-CHAVE-6=""/>
- <AREAS-DO-CONHECIMENTO>
    <AREA-DO-CONHECIMENTO-1 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS_EXATAS_E_DA_TERRA" NOME-DA-AREA-DO-CONHECIMENTO="Ciência da Computação"
    NOME-DA-SUB-AREA-DO-CONHECIMENTO="Recuperação de Informação" NOME-DA-ESPECIALIDADE=""/>
  </AREAS-DO-CONHECIMENTO>
```

Figura 5.2: Recorte das áreas de interesse

Já cadastrado no ambiente, o usuário tem acesso aos componentes de fichamento e o módulo de esclarecimento de dúvidas, ilustrado pela figura 5.3. Os modelos de fichamentos são norteados pela descrição apresentada no capítulo 2, como visto na figura 5.4 e 5.5.

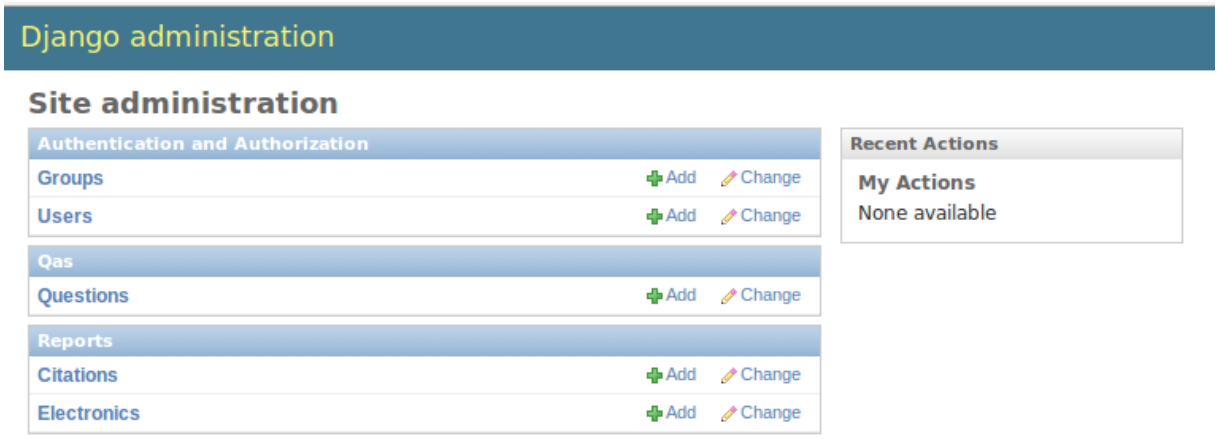

Figura 5.3: Interface de administração do recorte

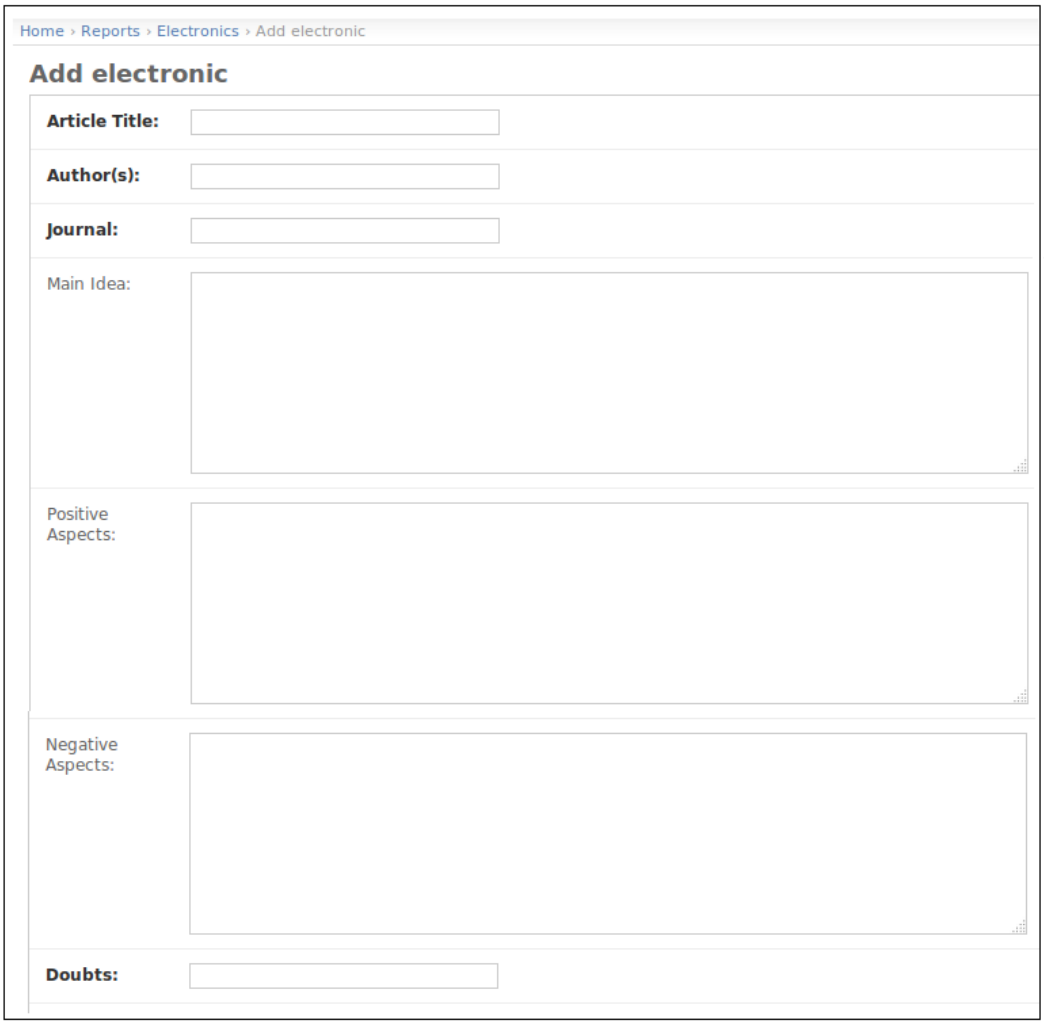

Figura 5.4: Recorte do fichamento do tipo eletrônico

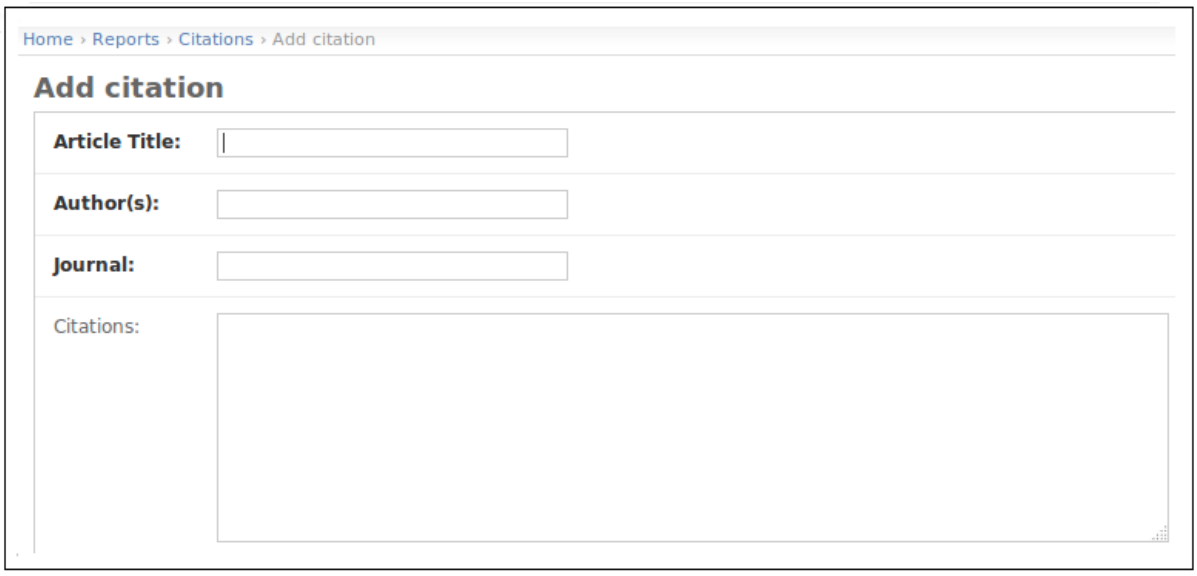

Figura 5.5: Exemplo de fichamento de citação

O perfil do usuário apresenta as informações mais relevantes a seu respeito, como as áreas de interesse, fichamentos, perguntas enviadas ao ambiente, seus grupos. Este perfil é atualizado de acordo com as interações do usuário com o sistema, o Django já possui um componente de histórico, onde grava todas as ações executadas no sistema.

Já o componente de QA do ambiente AICAPA apresenta o campo para que o usuário possa enviar a pergunta que, depois de enviada, inicia as atividades de processamento. Um exemplo de envio de pergunta está ilustrado na figura 5.6.

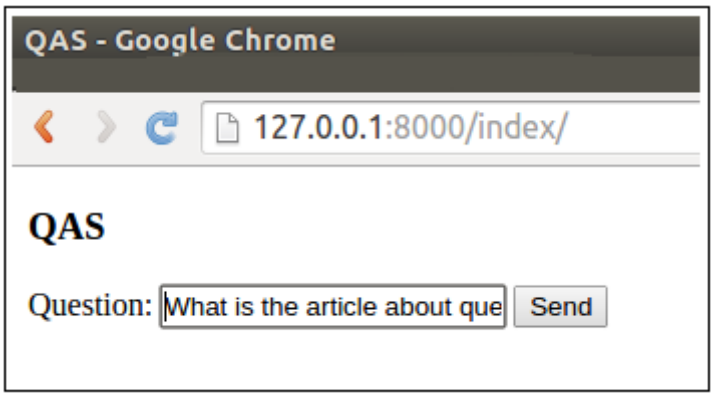

Figura 5.6: Exemplo de envio de pergunta

5.1.2 Processamento

Respeitando a arquitetura MVC do Django, os arquivos das aplicações implementadas seguem a mesma estrutura e nomenclatura. Nesse caso, as funções que correspondem aos módulos *query* e *search* são definidos nos arquivos *models.py*.

Para cada tipo de pergunta é definida uma função de análise, já que para cada tipo é esperado um tipo específico de resposta. Contudo, todas as entradas (perguntas) passam pelas mesmas tarefas de pré-processamento que são a normalização da pergunta, quebra da sentença em tokens e remoção das *stopwords.* O pseudocódigo descrito na figura 5.7 ilustra os passos percorridos pela pergunta na etapa de pré-processamento.

```
1   variáveis
2   pergunta, tokens, filtrada
3
4   início: análise da pergunta
5       leia pergunta
6       normatize pergunta
7       tokens = quebrar pergunta
8       filtrada = tokens - stopwords
9       retorne filtrada
10
11  fim
```

Figura 5.7: Pseudocódigo - Pré-processamento da pergunta

Para melhor compreensão e acompanhamento das ações do sistema, foi simulado um exemplo de execução com a entrada: *"What is the article about question answering system that has most citations?*".

O primeiro passo na execução do sistema é normalizar a entrada, ou seja, substituir caracteres de acentuação, converter toda a sentença em caracteres minúsculos e remover caracteres de pontuação. O próximo passo é quebrar a entrada em *tokens.* O NLTK dispõe de diversos métodos de tokenização, nesse caso optamos por utilizar a biblioteca *RegexpTokenize* e *word_tokenize*, que compara a entrada à uma expressão regular e divide a sentença em termos, como visto na figura 5.8.

```
>>> from nltk import word_tokenize, pos_tag
>>> question = "What is the article about question answering systems that has mo
st citations?"
>>> tokens = word_tokenize(question)
>>> tokens
['What', 'is', 'the', 'article', 'about', 'question', 'answering', 'systems', 't
hat', 'has', 'most', 'citations', '?']
>>> tokens
['What', 'is', 'the', 'article', 'about', 'question', 'answering', 'systems', 'that', 'has', 'most
', 'citations', '?']
>>>
```

Figura 5.8: Tokenização da sentença

Em seguida é realizada a remoção das *stopwords*, nesse caso são os termos: *is, the, about, that* e *has*. A eliminação destas palavras é realizada pelo componente do NLTK: *nltk.stopwords*. Esta biblioteca faz a comparação entre os *tokens* da sentença e uma lista de palavras na língua inglesa compiladas pelo NLTK (figura 5.9).

```
>>> stopwords = stopwords.words('english')
>>> question_clean = [w for w in tokens if w not in stopwords]
>>> question_clean
['What', 'article', 'question', 'answering', 'systems', 'citations', '?']
>>>
```

Figura 5.9: Remoção das stopwords

O passo seguinte é identificar as palavras-chave da sentença, neste caso são: *what*, *article*, *question answering system* e *citations*. Apesar de não ser uma palavra de posse de significado para sentença, o termo *citations* e o termo *article* servem de parâmetro para o *parser* que efetua a recuperação de informação a partir da busca realizada.

Todas essas atividades compõe a etapa de análise da pergunta, a saída dessa etapa é determinante para a busca e extração da resposta.

5.1.3 Núcleo do Sistema

Para construir a query de busca é necessário identificar o tipo de pergunta e, consequentemente, o tipo de resposta procurada. Para o exemplo utilizado procura-se o título do artigo sobre sistemas de esclarecimento de dúvidas com o maior número de citações. Na sentença existem três parâmetros para definição da query: o tipo de pergunta (qual), o foco (sistemas de esclarecimento de dúvidas) e um termo diferencial (número de citações). Nesse caso em o termo “citações” é o parâmetro para determinar a resposta esperada.

Para identificar os termos que vão compor a query de busca, é necessário realizar análises semânticas nos termos restantes. Nesse protótipo, faz-se uso de uma base de comparação, onde estão armazenados os termos de parâmetro que serão descartados. Estes termos serão referência para as próximas análises. A base de comparação tem como primeira entrada as marcações referentes a página de possíveis candidatos recuperada, os parâmetros iniciais nesse caso seriam os termos da estrutura dos artigos, por exemplo: url, *citations*, *author*, entre outros. A cada novo termo identificado é adicionado à base de comparação.

A cadeia de processos de análise e seleção das palavras-chave é descrita através do pseudocódigo da figura 5.10.

```
variáveis
filtradas, primeira lista, palavras-chave

arquivo
parâmetros

início: construção query
    primeira lista = filtradas - primeiro_termo
    parâmetros =+ primeiro_termo
    se (primeira_lista[termo] está em parâmetros) então:
        segunda_lista = primeira_lista - primeira_lista[termo]
        adicionar primeira_lista[termo] em parâmetros
        retorne segunda_lista
    caso contrário:
        retorne primeira_lista
fim
```

Figura 5.10: Pseudocódigo - Seleção das palavras-chave

Como pode ser observado no pseudocódigo, o primeiro termo dos tokens é removido, esse token corresponde ao tipo da pergunta (*what, who, when, list*), logo não vão fazer parte do conjunto de palavras-chave. Em seguida, verifica-se o restante dos tokens (primeira_lista) comparando cada token com os termos armazenados na base de comparação. Caso algum termo seja encontrado, é removido da lista e acrescentado à base, até que todos os tokens presentes na base sejam removidos. Os tokens restantes serão os termos que vão compor a query de consulta.

A busca no portal Google Acadêmico é realizada através do *parser*, para isso é necessário que a query atenda os parâmetros estabelecidos por este componente. A figura 5.11 ilustra os processos de definição da query.

```
querier = scholar.ScholarQuerier()
settings = scholar.ScholarSettings()
querier.apply_settings(settings)

query = scholar.SearchScholarQuery()
query.set_words(question)
```

Figura 5.11: Definição da query

As primeiras três linhas do código referem-se à configuração do *parser*, que são definidas pelo script do *parser* em si, necessitando apenas de definir as chamadas do script. Por se tratar de um script, o *parser* precisa de configurações básicas para ser chamado como função dentro de outra aplicação. As últimas linhas referem-se à atribuição da query, que recebe as palavras-chave selecionadas.

A query é então executada, com auxílio do *parser* e o resultado da busca, geralmente, é uma página HTML com todos os atributos, incluindo todas as tags de marcação próprias da linguagem, como visto na figura 5.12.

```
        <a href="https://scholar.google.com.br/scholar?cites=7823206262636285928&amp;as_sdt=2005&amp;sciodt=0,5&amp;hl=pt-BR">
         Citado por 165
        </a>
        <a href="https://scholar.google.com.br/scholar?q=related:6A8MkaWckWwJ:scholar.google.com/&amp;hl=pt-BR&amp;as_sdt=0,5">
         Artigos relacionados
        </a>
        <a class="gs_nph" href="https://scholar.google.com.br/scholar?cluster=7823206262636285928&amp;hl=pt-BR&amp;as_sdt=0,5">
         Todas as 3 versões
        </a>
        <a aria-controls="gs_citc" aria-haspopup="true" class="gs_nph" href="https://scholar.google.com.br/scholar?hl=pt-BR&amp;q=question+answering+system&amp;btnG=&amp;lr=#" onclick="return gs_ocit(event,'6A8MkaWckWwJ','6')" role="button">
         Citar
        </a>
        <span class="gs_nph">
         <a href="https://scholar.google.com.br/scholar?hl=pt-BR&amp;q=question+answering+system&amp;btnG=&amp;lr=#" id="gs_svl6" onclick="return gs_sva('6A8MkaWckWwJ','6')" title='Salvar este artigo em "Minha biblioteca" para que eu possa ler ou citar mais tarde.'>
          Salvar
         </a>
         <span class="gs_svm" id="gs_svo6">
```

Figura 5.12: Recorte da saída de recuperação do Google Acadêmico

Após a recuperação da página HTML, esta passa por nova análise do *parser*, com o intuito de remover das tags de formatação da página, transformando a página em um arquivo de texto.

Nessa etapa, a resposta é localizada no arquivo e extraída com auxílio do *parser* para Google Acadêmico e apresentada ao usuário, contendo informações de

acordo com a query gerada para consulta. Nesse caso, as informações recuperadas foram título, número de citações, ano de publicação, link do arquivo e trecho do abstract, como visto na figura 5.13.

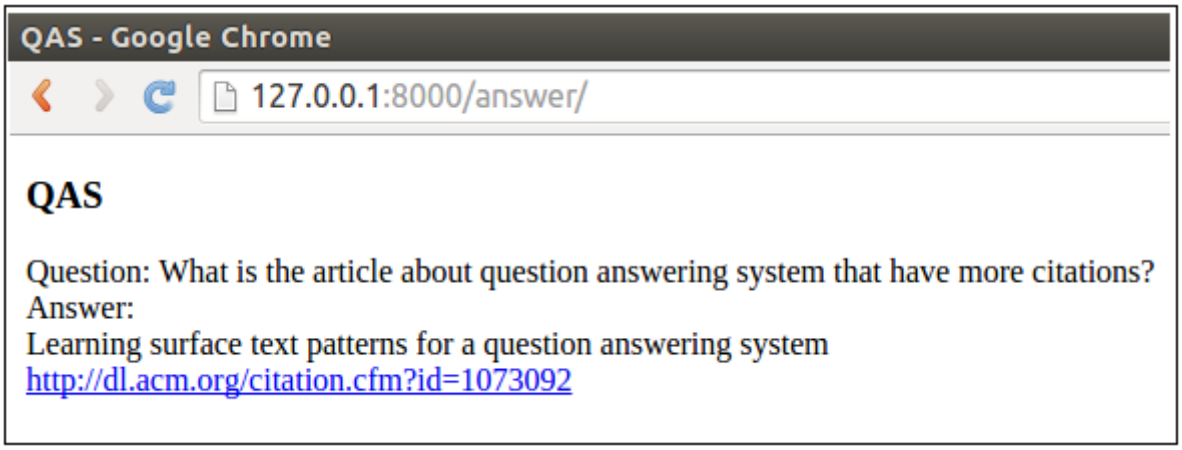

Figura 5.13: Exemplo de resposta gerada pelo Agente

A figura 5.14 mostra a saída do Agente para a pergunta "*List articles from author Harabagiu.*"

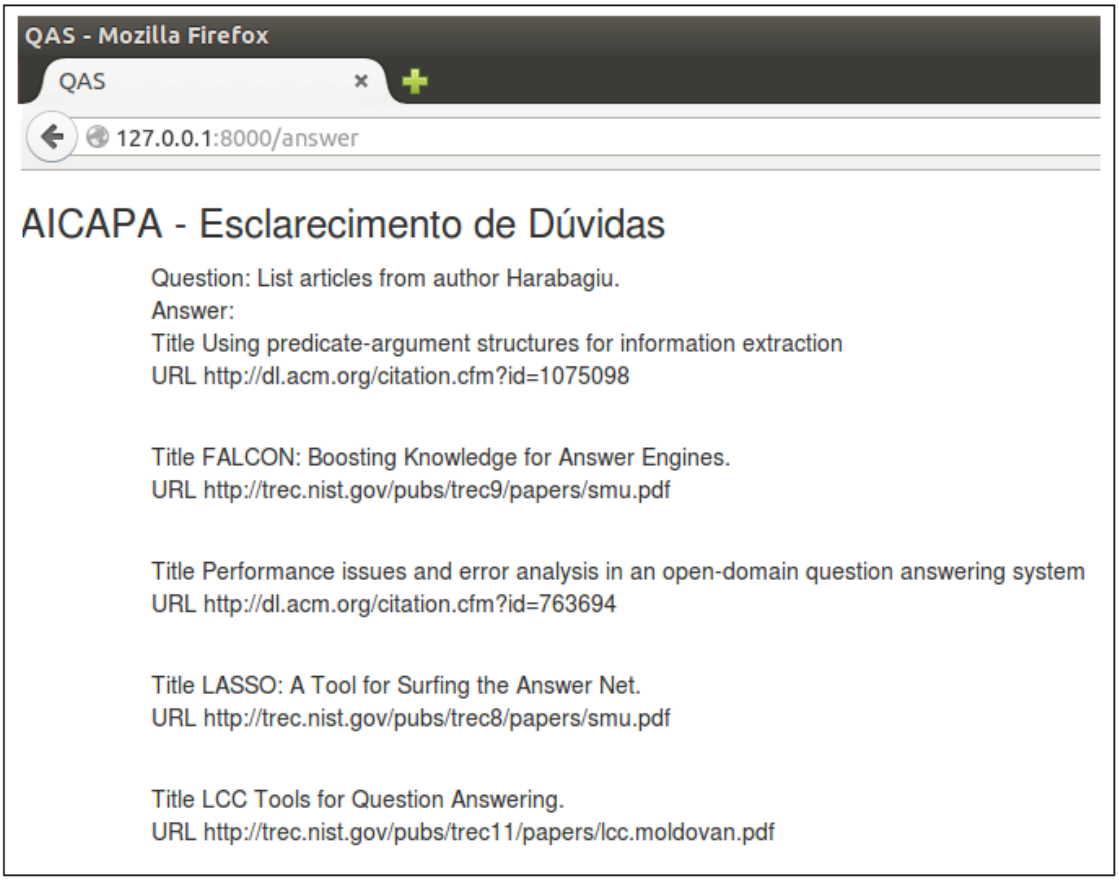

Figura 5.14: Saída do Agente

5.2 Testes e Resultados

Os testes foram realizados com um conjunto de 40 perguntas, variando entre perguntas, na língua inglesa, do tipo WH (*what, when, who*) e listas simples. As perguntas foram elaboradas com o intuito de abordar a etapa de levantamento de informações da uma produção acadêmica. O conjunto de perguntas utilizadas nos testes estão dispostas no apêndice A.

Para avaliar a qualidade das respostas apresentadas pelo Agente, foi utilizada a classificação proposta por MAGNINI et al. (2006). Serão, também, utilizadas métricas de avaliação comumente aplicadas a sistemas de recuperação de informação, que são:

- Abrangência (*recall*) – representa o número de perguntas respondidas pelo sistema dividido pelo número total de perguntas (POPESCU; ETZIONI; KAUTZ, 2003).
- Precisão (*precision*) – representa o número de respostas corretas dividido pelo número de perguntas respondidas (CIMIANO; HAASE; HEIZMAM, 2007).

Todas as respostas classificadas como incorretas, inexatas ou não suportadas serão encaminhadas aos especialistas respondedores. A tabela 5.1 apresenta os resultados em relação ao número de respostas corretas em relação ao número total de respostas, de acordo com a classificação proposta.

Tabela 5.1: Abrangência do Agente de Dúvidas

Classificação	Número de respostas	Porcentagem
Correta	16	40%
Inexata	9	22,5%
Não suportada	14	35%
Incorreta	1	2,5%
Total	40	100%

A abrangência do sistema, de acordo com a definição proposta, foi de 100%, já que todas as perguntas inseridas foram respondidas, mesmo que de forma não adequada.

Do conjunto de 40 perguntas a divisão do tipo de pergunta foi a seguinte:

- Perguntas do tipo *what:* 15;
- Perguntas do tipo *when*: 4;
- Perguntas do tipo *who*: 3;
- Perguntas do tipo lista: 18.

A tabela 5.2 apresenta o número de respostas corretas de cada tipo de pergunta.

Tabela 5.2: Número de respostas corretas para cada tipo de pergunta

Tipo de pergunta	Número de respostas	Porcentagem
What	5	33,33%
When	2	50%
Who	0	0%
Lista	9	50%

Analisando os resultados obtidos, podemos extrair as seguintes conclusões:

- Restringir a pergunta a modelos padronizados podem interferir na qualidade da resposta. Algumas das perguntas elaboradas envolveram raciocínio subjetivo, que o modelo aplicado para resolução das perguntas não foi capaz de interpretar;
- Perguntas que se adequam as informações mais concisas, como ano de publicação e número de citações, mostram ser mais eficientes para o modelo aplicado;
- Perguntas do tipo *who* obtiveram os piores resultados, já que o tipo esperado de resposta era o nome de uma ou mais pessoas ou entidades;
- Respostas classificadas como inexatas apresentaram informações além do necessário para a pergunta em questão, ou seja, cabe ao usuário a decisão se a resposta foi adequada ao seu questionamento;

- O expressivo número de perguntas não suportadas está diretamente ligada ao tipo de resposta esperado da pergunta, em sua maioria necessitam de conceitos descritivos, o que é naturalmente não suportado pela fonte de dados externos utilizadas nos testes (Google Acadêmico).

6 CONSIDERAÇÕES FINAIS

Este trabalho apresentou a proposta de um Sistema de Perguntas e Respostas, parte do conjunto de funcionalidades de um ambiente colaborativo em favor da produção acadêmica. No desenvolvimento foram utilizadas tecnologias livres, sendo adaptadas para o propósito descrito. Os atuais sistemas de perguntas são utilizados de forma individual, sem qualquer interação com outras ferramentas de cunho acadêmico. O número de Sistemas de Perguntas e Respostas disponíveis atualmente corrobora a necessidade de sistemas customizados à necessidade do usuário.

Através dos experimentos realizados em perguntas relacionadas ao levantamento de artigos para a pesquisa acadêmica, foi constatado que a proposta de um ambiente integrado e colaborativo é viável e importante para o meio acadêmico. Contudo, a aplicação de mais de uma técnica de raciocínio semântico mostra-se necessária para a melhor resolução das perguntas. Ainda assim, é importante destacar que o papel dos componentes humanos é de fundamental relevância para a determinação da qualidade das respostas e difusão de conhecimento.

6.1 Lições Aprendidas

Durante a concepção e desenvolvimento da solução proposta neste trabalho observou-se como a Inteligência Artificial, em especial a área de Processamento de Linguagem Natural, pode contribuir para o campo de pesquisa e produção acadêmica.

Com a metodologia aplicada foi possível observar que a utilização de diferentes métodos, indo de técnicas de análise e inferências semânticas até a recuperação de informação de bases textuais locais.

A utilização de técnicas de Recuperação de Informação foi um meio viável para atingir o objetivo proposto, contudo percebeu-se que a união de outras técnicas de Inteligência Artificial podem contribuir grandemente para a melhoria da qualidade das respostas obtidas.

A concepção da proposta apresentada, principalmente as escolhas e definição da abordagem adotada, foi de fundamental importância para o amadurecimento pessoal, em especial à tomada de decisão e respectiva justificativa.

6.2 Limitações do Trabalho

O sistema descrito neste trabalho apresenta limitações, que são as que segue:

- Língua: o protótipo foi desenvolvido para responder, somente, perguntas na língua inglesa. Essa restrição foi imposta por conta das bibliotecas de análise léxica e semânticas utilizadas.
- Perguntas: o protótipo apresenta respostas para um grupo de perguntas factoides e listas. A abordagem adotada na implementação do sistema utiliza as palavras-chave extraídas da pergunta original para criar uma query que possa ser executada no portal de busca Google Acadêmico. Perguntas subjetivas não apresentam resultado significante nesse portal.
- Base de conhecimento: optou-se por usar o Google Acadêmico como base de conhecimento, por ser diretamente utilizado como fonte de pesquisa para publicações acadêmicas. Essa restrição exclui a utilização de outros formatos de mídia e a perguntas subjetivas ou que necessitem de resposta do tipo definição.
- Conexão à Internet: por fazer uso de uma ferramenta de busca online, este protótipo necessita a disponibilidade de conexão à Internet.

6.3 Trabalhos Futuros

Tendo em vista a gama de possíveis abordagens para implementação e/ou customização de Sistemas de Perguntas e Respostas, a proposta apresentada representa uma pequena parcela das possibilidades a serem exploradas. Como possíveis trabalhos futuros citamos:

- Ampliar as bases de conhecimento do sistema, ou seja, possibilitar que o Agente de Busca e Recuperação atue em diferentes mecanismos de busca, inclusive APIs disponibilizadas por outros portais acadêmicos, como o IEEE;
- Expandir o tipo de perguntas resolvidas pelo Agente de Dúvidas, ou seja, permitir que o sistema interaja com perguntas de outros tipos semânticos, principalmente perguntas que necessitam de conceitos descritivos;
- Possibilitar que o Agente de Dúvidas identifique, automaticamente, as dúvidas listadas nos fichamentos e possa buscar por candidatos à resposta diretamente

dos arquivos armazenados no Banco de Artigos, e também fazer uso dos grandes motores de busca, como o Google;

- Implementar modelos de perfil de usuário, para que o funcionamento tanto do Agente de Dúvidas quanto das funcionalidades do AICAPA, sejam customizados de acordo com a modelagem de cada tipo de perfil;
- Aplicar técnicas de análise semântica, por exemplo. O uso de mais técnicas de interpretação semântica pode aumentar a precisão das respostas extraídas pelo sistema.

REFERÊNCIAS BIBLIOGRÁFICAS

AMORIM, M. T. C. F.; CURY, D.; MENEZES, C. S. Um Sistema Inteligente Baseado em Ontologia para Apoio de Esclarecimento de Dúvidas. **Anais do XXII SBIE - XVII WIE**, 2011.

ANDRADE, J. C. et al. QSabe - Um Ambiente Inteligente para Endereçamento de Perguntas em uma Comunidade Virtual de Esclarecimento. **Proceedings of the First Latin American Web Congress**, 2003.

BIRD, S. **NLTK: The Natural Languague Toolkit**COLING/ACL 2006 Interactive Presentation Sessions. **Anais**...Sydney: jul. 2006

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Languague Processing with Python**. California: O'Reilly, 2009.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer Networks**, v. 30, n. 1-7, p. 107–111, 1998.

BURGER, J. et al. **Issues, tasks and program structures to roadmap research in question & answering (Q&A)**Document Understanding Conferences Roadmapping Documents. **Anais**...2001

CARBONELL, J. et al. Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization. **Q&A Summarization Vision**, 2000.

CARPINETO, C.; ROMANO, G. A Survey of Automatic Query Expansion in Information Retrieval. **ACM Computer Survey**, v. 44, n. 1, p. 50, 2012.

CIMIANO, P.; HAASE, P.; HEIZMAM, J. Porting Natural Language Interfaces between Domains – An Experimental User Study with the ORAKEL System –. **In Proceedings of the 12th international conference on Intelligent user interfaces**, p. 180–189, 2007.

DAMLJANOVIC, D.; AGATONOVIC, M.; CUNNINGHAM, H. **FREyA: An interactive way of querying Linked Data using natural language**The Semantic Web: ESWC 2011 Workshops. **Anais**...2012

DIAS, P. **Comunidades de Conhecimento e Aprendizagem Colaborativa**, jul. 2001. Disponível em: <http://www.prof2000.pt/users/mfflores/teorica6_02.htm>

GREEN JR., B. F. et al. **Baseball: An Automatic Question-answerer**Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference. **Anais**...: IRE-AIEE-ACM '61 (Western).New York, NY, USA: ACM, 1961Disponível em: <http://doi.acm.org/10.1145/1460690.1460714>

GUO, Q.; ZHANG, M. Question Answering System Based on Ontology and Semantic Web. **RSKT 2008 : the 3rd International Conference on Rough Sets and Knowledge Technology**, p. 652–659, 2008.

GUPTA, P.; GUPTA, V. A Survey of Text Question Answering Techniques. **International Journal of Computer Applications**, 2012.

HARABAGIU, S. M. et al. Answering complex, list and context questions with LCC's Question-Answering Server. 2001.

HIRSCHMAN, L.; GAIZAUSKAS, R. Natural Language Question Answering: the view from here. **Natural Language Engineering**, 2001.

HOLOVATY, A.; KAPLAN-MOSS, J. **The Definitive Guide to Django: Web Development Done Right**. [s.l.] Apress, 2009.

KANGAVARI, M. R.; GHANDCHI, S.; GOLPOUR, M. Information Retrieval: Improving Question Answering Systems by Query Reformulation and Answer Validation. **World Academy of Science, Engineering and Technology**, 2008.

KAUARK, F. S.; MANHÃES, F. C.; MEDEIROS, C. H. **Metodologia da Pesquisa: um guia prático**. [s.l.] Editora Via Litterarum, 2010.

KAUFMANN, E.; BERNSTEIN, A. **How useful are natural language interfaces to the semantic web for casual end-users?**. [s.l.] Springer, 2007.

KREIBICH, C. **scholar.py - A Parser for Google Scholar**, 2013. Disponível em: <http://icir.org/christian/scholar.html>

LAKATOS, E. M.; MARCONI, M. D. A. Fundamentos de Metodologia Científica. In: [s.l.] Editora Atlas S.A., 2003. p. 44–73.

LARSEN, P. O.; VON INS, M. The rate of growth in scientific publication and the decline in coverage provided by science citation index. **Scientometrics**, v. 84, p. 575–603, 2010.

LEHNERT, W. Human and Computational Question Answering*. **Cognitive Science**, v. 1, n. 1, p. 47–73, 1977.

LOPEZ, V. et al. AquaLog: An ontology-driven question answering system for organizational semantic intranets. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 5, n. 2, p. 72–105, 2007.

MAGNINI, B. et al. **Overview of the CLEF 2006 Multilingual Question Answering Track**CLEF 2006 Conference. **Anais**...2006

MANNING, C. D.; RAGHAVAN, P. **An Introduction to Information Retrieval Online**, 2009. Disponível em: <http://dspace.cusat.ac.in/dspace/handle/123456789/2538>

MAYBURY, M. T. Toward a question answering roadmap. **New Directions in Question Answering**, v. 2003, p. 8–11, 2003.

MAYBURY, M. T. New directions in question answering. 2004.

MEDEIROS, J. B. **Redação Científica: a prática de fichamentos, resumos, resenhas**. [s.l.] Editora Atlas S.A., 2006.

MENEZES, C. S.; TAVARES, O. L.; PESSOA, J. M. QSabe - Trocando Experiências sobre Informática Educativa em uma Rede de Educadores. **Revista Brasileira de Informática na Educação**, 1998.

MOLDOVAN, D. et al. Performance Issues and Erros Analysis in an Open-Domain Question Answering System. **ACM Transactions on Information Systems**, v. 21, n. 2, p. 133–154, abr. 2003.

MOLDOVAN, D. I. et al. **LASSO: A Tool for Surfing the Answer Net.**TREC. **Anais**...1999

NLTK. **NLTK Project**, 2013. Disponível em: <http://www.nltk.org>

PESSOA, J. M. **Desenvolvimento Orientado a Agentes: Uma Experiência com Agentes de Interface**. [s.l.] UFES, 1997.

POPESCU, A.-M.; ETZIONI, O.; KAUTZ, H. Towards a theory of natural language interfaces to databases. **Proceedings of the 8th international conference on Intelligent user interfaces IUI 03**, v. terfaces, p. 149–157, 2003.

REDDY, S.; SCIENCE, I.; STATE, K. Use of Information Sources by Research Scholars : A Case Study of Gulbarga University. **Library**, v. 2010, p. 1–4, 2010.

ROSSUM, G. VAN. **Python Reference Manual Release 2.5.2**, 2008.

RUSSELL, K. **Review of NLTK 1.2 (Linguist List)**. [s.l: s.n.]. Disponível em: <http://linguistlist.org/issues/14/14-3165.html>.

SAIAS, J. M. G. **Contextualização e Activação Semântica na Seleção de Resultados em Sistemas de Pergunta-Resposta**. [s.l.] Universidade de Ávora, 2010.

SENADO. **Produção Científica no Brasil: um salto no número de publicações** senado.gov.br, 2010. Disponível em: <http://www.senado.gov.br/noticias/Jornal/emdiscussao/inovacao/investimento-inovacao-tecnologica-finep-pesquisadores-brasil/producao-cientifica-no-brasil-um-salto-no-numero-de-publicacoes.aspx>

SIMMONS, R. F. Answering Questions by Computer: A Survey. **Communications of the ACM**, 1965.

STRZALKOWSKI, T.; HARABAGIU, S. **Advances In Open Domain Question Answering**. [s.l.] Springer, 2008.

THE ECONOMIST. **All too much**, 2010. Disponível em: <http://www.economist.com/node/15557421>

VICEDO, J. L.; MOLLÁ, D. Open-Domain Question-Answering Technology: State of Art and Future Trends. **ACM Journal Name**, 2001.

WEBBER, B.; WEBB, N. Question Answering. In: CLARK, A.; FOX, C.; LAPPIN, S. (Eds.). . **The Handbook of Computational Linguistics and Natural Language Processing**. [s.l.] Wiley-Blackwell, 2010. p. 630–654.

ZADEH, L. A. Fuzzy Logic and the Semantic Web. In: SANCHEZ, E. (Ed.). . [s.l.] Elsevier B. V., 2006. p. 163–210.

ZHENG, Z. **AnswerBus question answering system.** Proceedings of the second international conference on Human Language Technology Research. **Anais**...2002

APÊNDICE A

Lista de Perguntas

1. What is the first article published about Question and Answering System?
2. When was published the first article about Question Answering System?
3. List the complete reference of article X.
4. What is question answering system?
5. What is the difference between QAS and search systems?
6. Who is the author that have more citations on QAS?
7. What is the most relevant article about QAS?
8. List the articles about QAS that have more citations.
9. List sources of academic research.
10. Who is the author of article X?
11. What is the title of article from author X?
12. List the academic fields related to QAS.
13. What are the authors that have more citations on QAS?
14. What are the journals that have more publications on Computer Science?
15. What is Natural Language Processing?
16. List some examples of Question Answering Systems.
17. List articles from author X.
18. When was article X published?
19. What are the first articles about X and Y?
20. List articles published in year X.
21. Who are the authors of article X?
22. What is the main idea of article X?
23. List articles related to article X.
24. List authors that have articles about X.
25. List articles from author X and Y.
26. List articles about X and Y.
27. List articles about X with Z methods.
28. List articles about X using N technologies.
29. Who is the author that wrote about X.
30. List authors that wrote about X.
31. List the most recent articles about X.
32. What is the most recent article that X wrote?
33. List articles published after year X.
34. List articles from author X after year X.
35. What are the articles from author X without Y.
36. What is the last article published about X?
37. When was published the last article about X?
38. When author X publish article Y?
39. What is the number of citations on article X?
40. List articles about Y with the most number of citations after year X.

GLOSSÁRIO

AIML – linguagem baseada em XML, desenvolvida para criar diálogos semelhantes a linguagem natural.

API – é um conjunto de rotinas e padrões estabelecidos por um software para utilização de suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação do software, mas apenas utilizar seus serviços.

Corpora – plural de *corpus,* são grandes bases estruturadas de texto.

Framework – conjunto de classes que colaboram para realizar uma responsabilidade para um domínio de um subsistema da aplicação.

Ontologia – modelo de dados que representa um conjunto de conceitos dentro de um domínio e os relacionamentos entre este.

Parser – analisador de strings de símbolos, seja em linguagem natural ou computacional, conforme as regras de uma gramática formal.

POS Tagging – Um *part-of-speech tagger* (*tagging)* é um software que lê um texto em algum idioma e atribui partes da linguagem para cada palavra.

Respostas Factoides – respostas que apresentam trechos de texto curtos, diretos e factuais.

RDF – modelo padrão para troca de dados na Web.

SPARQL – linguagem de consulta para RDF.

Stemming – processo de redução de um termo à sua raiz comum.

Stopwords – palavras que não carregam nenhum valor semântico à frase, sendo filtradas antes ou depois do processamento de linguagem natural.

Token – é um conjunto de caracteres com significado coletivo.

WordNet – base léxica na língua Inglesa.