UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Recuperação Contextual de Informação na *Web* a partir da Análise de Mensagens e Enriquecimento em Dados Abertos: Explorando o contexto de ensino/aprendizagem

Eduardo Fritzen

Orientadores

Leila Cristina Vasconcelos de Andrade

Sean Wolfgand Matsui Siqueira

RIO DE JANEIRO, RJ – BRASIL

Junho de 2012

RECUPERAÇÃO CONTEXTUAL DE INFORMAÇÃO NA *WEB* A PARTIR DA ANÁLISE DE MENSAGENS E ENRIQUECIMENTO EM DADOS ABERTOS: EXPLORANDO O CONTEXTO DE ENSINO/APRENDIZAGEM

Eduardo Fritzen

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

_____
Leila Cristina Vasconcelos de Andrade, D.Sc. – UNIRIO

_____
Sean Wolfgand Matsui Siqueira, D.Sc. – UNIRIO

_____
Flávia Maria Santoro, D.Sc. – UNIRIO

_____
Ig Ibert Bittencourt Santana Pinto, D.Sc. – UFAL

RIO DE JANEIRO, RJ – BRASIL

JUNHO DE 2012

Aos meus pais José e Linda

Agradecimentos

A Deus pelo dom da vida.

Aos meus pais que com toda simplicidade educaram, mostrando-me que sonho e luta são essenciais para a conquista.

A minha irmã, Carolina, que, apesar da distância, sempre torceu por mim.

Ao meu amor Marta Teixeira, que soube compreender minhas ausências nos finais de semana e feriados por estar estudando. Obrigado pela paciência e parceria.

A todos familiares e amigos, pelo apoio e companheirismo que vivenciamos.

A professora Leila pela orientação e ensinamentos, mas acima de tudo pela força e doses de motivação que injetou durante os momentos mais difíceis, fundamentais para perseverança desta caminhada. Para mim, uma "mãezona".

Ao professor Sean pela orientação e ensinamentos. Competente, sempre com dicas oportunas na manga, presente, paciente e disposto a ajudar. Para mim, um amigo.

Aos professores Ângelo, Flávia, Simone, Fernanda, Kate e Lucena que contribuíram para o meu aprendizado e formação. Muito obrigado pela amizade e profissionalismo.

Aos colegas Prates, Krejci, Polo, Patrícia, Débora, Edvaldo, Marcos, Viviane, Mateus, Elberth, Fabiano pelos inúmeros auxílios e contribuições, em diferentes momentos, sempre prestativos.

Aos colegas da DATAPREV, Verônica, Durvalino, Cláudio, Guilherme, Bianca e Adilson, por compreender minhas ausências durante a jornada de trabalho. Aos gerentes de departamento Neiva e Nelson, por viabilizar essa idéia.

A todos os alunos do curso de bacharelado em sistemas de informação da UNIRIO (turmas FSI 2011.1 e FSI 2011.2) que participaram no estudo de caso.

Fritzen, Eduardo. **Recuperação Contextual de Informação na *Web* a partir da Análise de Mensagens e Enriquecimento em Dados Abertos: Explorando o contexto de ensino/aprendizagem.** UNIRIO, 2012. 190 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Os motores de busca transformaram o processo de se encontrar algo na Web, enquanto que as plataformas sociais revolucionaram as formas de comunicação e relacionamento entre as pessoas. A popularidade destes sistemas de informação influencia diretamente o cotidiano das pessoas. Entretanto, em geral, as plataformas que possibilitam a formação de redes sociais online fornecem apenas buscas nas próprias redes sociais ou disponibilizam motores de busca na Web tradicionais sem considerar o contexto destas redes sociais. Quando estas plataformas são utilizadas em ambientes de trabalho ou para apoiar processos de ensino/aprendizagem, considera-se que a relevância na recuperação de documentos na Web possa ser primordial nestas atividades. Para melhorar e contextualizar a recuperação de documentos na Web, este trabalho propõe: (i) a modelagem do contexto a partir da extração das mensagens em grupos de rede social e (ii) uso do contexto para melhorar a relevância na recuperação de documentos na Web. Técnicas de processamento de linguagem natural e recuperação da informação são utilizadas para a extração de termos relevantes das mensagens, que são enriquecidos a partir de dados abertos com apoio de tecnologias semânticas. De modo a possibilitar o processamento contínuo de mensagens, tecnologias de agentes de software também foram consideradas. Dois estudos de caso demonstraram que a captura do contexto enriquecido usando mensagens de discussão pode melhorar a relevância dos resultados das buscas na Web e contribuir com as discussões.

Palavras-chave: Recuperação de Informação, Extração de Informação, Contexto, Dados Abertos, Redes Sociais, Processamento de Linguagem Natural, Agentes, Aprendizagem Colaborativa, Sistemas de Informação.

ABSTRACT

The search engines have changed the process of finding things on the *Web*, while social *platforms* have revolutionized the way people communicate and relate to each other. The popularity of these information systems directly influences people's daily lives. However, in general, *platforms* for online social networks provide searches only over their own data (the social networks) or use search engines available on the *Web* without considering the context of these social networks. When these *platforms* are used in workplaces or to support teaching / learning processes, the relevance of retrieving documents on the *Web* could be essential to these activities. In order to enhance and contextualize the document retrieval on the *Web*, this work proposes: (i) modeling the context from the extraction of messages from social network groups, and (ii) using the context to improve the relevance of documents retrieved from the *Web*. Natural language processing and information retrieval techniques are used to extract relevant terms of the messages. Then, these terms are enriched from open data with the support of semantic technologies. In order to provide continuous processing of messages, *software* agent technologies were also considered. Two case studies showed that capturing the context of discussion messages using dynamic collaborative learning discussions can improve the relevance of search results on the *Web*.

Keywords: Information Retrieval, Information Extraction, Context, Open Data, Social Networks, Natural Language Processing, Agents, Collaborative Learning, Information Systems.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ACL	*Agent Communication Language*
API	Application Programming Interface
ASL	Análise Semântica Latente
CCS	*Collaborative Context Search*
CCSA	*Collaborative Context-Search Agent*
CSV	Comma-Separated Values
F-EXT-WS	*Web Service* para Processamento de Linguagem Natural
FIPA	*Foundation for Intelligent Physical Agents*
JSON	*JavaScript Object Notation*
KQML	*Knowledge Query Manipulation Language*
LEARN	*Algorithm Engineering and Machine Learning Laboratory*
LOD	*Linking Open Data*
LSA	*Latent Semantic Analysis*
NLP	Natural Language Processing
OMG	*Object Management Group*
OWL	*Web Ontology Language*
PLN	Processamento de Linguagem Natural
POST	*Part-Of-Speech Tagging*
POS	*Part-Of-Speech*
RDF	*Resource Description Framework*
RDFS	*Resource Description Framework Schema*
RI	Recuperação de Informação
RRI	Indexação Aleatória Reflexiva

SMA Sistema Multiagentes

SPARQL *SPARQL Protocol and RDF Query Language*

SRI Sistema de Recuperação de Informação

TF *Term Frequency*

TF-IDF *Term Frequency – Inverse Document Frequency*

TREC Text REtrieval Conference

URI Identificador Uniforme de Recursos (Uniform Resource Identifier)

URL Uniform Resource Locator (Localizador-Padrão de Recursos)

VSM *Vector Space Model*

XML *Extensible Markup Language*

W3C *World Wide Web Consortium*

WSDL *Web* Services Description Language

Capítulo 1 – Introdução

Este capítulo fornece uma visão geral da dissertação, bem como a motivação para o uso das discussões para contextualizar a recuperação de documentos na *Web*. A proposta de solução para o problema exposto é brevemente descrita, apresentando as linhas de ação que norteiam esta dissertação. A hipótese e a metodologia usada para testá-la também são apresentadas.

1.1. Motivação

Os novos recursos e os avanços das aplicações disponíveis na *Web*[1] influenciam cada vez mais o cotidiano das pessoas (BCG, 2012). Essas novas aplicações, chamadas de mídias sociais[2], são cunhadas sobre as tecnologias e ideologias da *Web* Social[3]. As mídias sociais fornecem espaço de colaboração e compartilhamento de conteúdo (como texto, imagem, som e vídeo), o que habilita o uso dessas plataformas sociais para intercâmbio de experiências em meio eletrônico.

As plataformas sociais possibilitam a formação de redes sociais online e merecem destaque pela sua popularidade, ascendente tanto em número de usuários quanto ao tempo que esses dedicam à navegação. Dos brasileiros que acessam a internet, 72% já

[1] Web, WWW ou World Wide Web são sinônimos e representam a porção hipertextual da Internet.
[2] Segundo Kaplan e Haenlein (2010), Mídia Social é um grupo de aplicações baseadas na Web que permitem a criação e a troca de conteúdo digital gerado pelo usuário.
[3] Segundo Porter (2008), Web Social (ou Web 2.0) é um conjunto de relações entre pessoas na Web suportadas por aplicações projetadas para apoiar e fomentar a interação social.

incorporaram às suas rotinas o hábito de navegar em pelo menos um sítio de rede social online (IBOPE, 2011).

A *Web* é uma gigantesca fonte de informação, que vem crescendo de forma muito acelerada nos últimos anos. Segundo estimativas da Netcraft (2011), existem mais de 500 milhões de sítios no mundo. Assim, outra tecnologia amplamente difundida e utilizada na navegação dos usuários corresponde aos motores de busca na *Web* (ALEXA, 2012). Motores de busca na *Web*, tais como Google[4] e Bing[5], foram concebidos para possibilitar a recuperação da informação desejada em meio às dezenas de milhões de páginas que compõem seus índices. Normalmente, estes motores de busca consideram dados estatísticos, históricos, popularidade do sítio e localização geográfica para priorização dos resultados (GOOGLE, 2012).

Ao observar o resultado das buscas realizadas nos motores de buscas disponibilizados nos dois sítios de redes sociais mais populares do Brasil, *Facebook*[6] e *Orkut*[7] (COMSCORE, 2011), percebe-se que o *Orkut* restringe os resultados da busca a elementos de sua própria rede, como pessoas e comunidades, enquanto no *Facebook* os resultados da *Web* são idênticos aos obtidos com o sítio de seu motor de busca *Web*, independente se a pesquisa originou-se no sítio do buscador ou a partir de um grupo de discussão no sítio da rede social.

No exemplo da Figura 1, a expressão de busca "ator especialização" retornou o mesmo conjunto de resultados se executada pelo sítio da rede social ou diretamente pelo sítio do referido motor de busca. Todos os dez primeiros resultados relacionam-se a sítios do domínio de "artes cênicas", embora o interesse de pesquisa seja especialização de atores no domínio de "modelagem de sistemas". Portanto, a priorização dos

[4] http://www.google.com
[5] http://www.bing.com
[6] http://www.facebook.com/
[7] http://www.orkut.com.br/

resultados não considera o contexto das redes sociais em que os motores de busca estão inseridos.

Figura 1 – Busca no sítio da Rede Social *Facebook*

A necessidade de usuários adquirirem documentos a partir da *Web* que os apoiem a participar e entender os assuntos discutidos em suas redes sociais foi o que motivou este trabalho.

1.2. Problema

A recuperação de informação desejada a partir da *Web* não é uma tarefa trivial (KOBAYASHI e TAKEDA, 2000). Os resultados dos motores de busca na *Web* atual não necessariamente refletem o que se busca em determinado momento. Por exemplo, ao se fazer uma busca na *Web* por uma palavra como “Flamengo”, pode-se querer obter resultados sobre um time de futebol, um bairro da cidade do Rio de Janeiro, um trecho da música do Djavan ou uma região ao norte da Bélgica. Assim, a priorização de resultados é um problema que deve ser investigado a fim de buscar meios de classificação que melhor se enquadrem às necessidades de informação do usuário.

Este problema se torna mais crítico quando as redes sociais envolvem atividades de trabalho ou de ensino/aprendizagem, pois as pesquisas na *Web* realizadas a partir

destes ambientes visam auxiliar o desenvolvimento dessas atividades. Melhorar a precisão dos resultados das buscas pode estimular as discussões no grupo e promover a colaboração. Por exemplo, em um ambiente educacional, a relevância dos documentos obtidos a partir da *Web* pode apoiar o aprendizado, uma vez que o conteúdo retornado será possivelmente mais adequado à necessidade de informação do aprendiz.

Assim, o fato dos resultados de busca de documentos na *Web* em plataformas de redes sociais atualmente não serem precisos o suficiente para apoiar atividades de ensino/aprendizagem nas próprias redes é o problema a ser resolvido nesta pesquisa.

1.3. Hipótese

SE a consulta do usuário em motores de busca na *Web* a partir de plataformas de redes sociais for expandida com termos extraídos das mensagens destas redes e enriquecidos com dados abertos[8], ENTÃO os resultados da busca expandida trarão documentos mais relevantes que os obtidos com a consulta original.

A hipótese é falseável, pois os resultados podem ser piores que os obtidos com a consulta original, por uma série de fatores, tais como: (i) a obtenção de termos extraídos das mensagens das redes sociais pode não ser significativa o suficiente para uma expansão de consulta que modele o contexto das discussões, (ii) o enriquecimento dos termos pode não ser possível ou mesmo relevante para a expansão das consultas; (iii) os motores de busca podem estar preparados para responder consultas com poucas palavras-chave e portanto a expansão de consulta pode resultar em documentos menos relevantes, (iv) as palavras-chave informadas pelo usuário podem ser suficientes e retornar documentos mais relevantes do que uma possível expansão e (v) a sugestão de termos para expansão pode não ser significativa.

[8] Dados abertos devem ser livres, disponíveis na Web e sem restrições de direitos autorais.

1.4. Enfoque de solução

O enfoque de solução usa mensagens de *softwares* de comunicação, como as discussões em grupos das redes sociais para gerar o contexto e usá-lo para extrair e sugerir palavras-chave que mais se aproximem da expressão de busca informada pelo usuário. Estas palavras são enriquecidas a partir de dados abertos e poderão ser combinadas pelo usuário e adicionadas à consulta original. Pode-se dizer então que a consulta é expandida a partir da construção de vocabulários gerados automaticamente por algoritmos de processamento de texto, com a ajuda do usuário.

Em geral, a expressão de busca (conjunto de palavras-chave) informada pelo usuário é composta por poucos termos e, portanto, pouco representativa do contexto do domínio do usuário e suscetível a ambiguidades que degradam o processamento e resultado da busca (ou seja, os documentos que compõem o resultado não são tão relevantes para o usuário de acordo com seu contexto). Segundo levantamento feito pela empresa Experian Hitwise, entre 26/08/2011 e 26/11/2011, 66,55% das buscas utilizam no máximo três palavras (o número de buscas contendo apenas uma palavra foi de 26,03%, seguidas por buscas de duas palavras, com 20,63%, e três palavras, 19,89%). Pesquisas maiores, com oito (8) ou mais palavras totalizaram apenas 4,55% das expressões de busca (EXPERIAN, 2011). Outra informação relevante sobre o perfil das buscas na *Web* é que apenas os cinco (5) primeiros resultados são efetivamente acessados pelos usuários (SPINK e JANSEN, 2004), o que reforça a importância da priorização dos resultados.

Os usuários, em geral, têm dificuldades em formular essa expressão de busca. Adicionar mais termos à consulta, e, quiçá, expressões lógicas conjuntivas, é uma alternativa para melhorar a precisão dos resultados, visto que, em geral, palavras isoladas geram alto nível de ambiguidade. Além da quantidade, a qualidade dos termos

também deve ser considerada, frente à dificuldade que grande parte dos usuários tem em definir quais palavras-chave são boas representantes para os documentos de seu interesse e que deverão compor a sua expressão de busca (FERNEDA, 2003). No exemplo, caso se procurasse pelo bairro do flamengo, poder-se-ia usar a expressão "flamengo AND bairro AND 'rio de janeiro'", para melhorar os resultados da busca. Entretanto, menos de 5% das buscas na *Web* usam recursos de busca avançados, como sequência de duas ou mais palavras contíguas (por exemplo, 'rio de janeiro' usado entre aspas) ou operadores lógicos (AND, OR e NOT) (SPINK *et* al., 2001).

Para contornar esta questão, em geral, usa-se um modelo de domínio para fornecer o contexto da consulta, seja a partir de uma modelagem feita por especialistas ou a partir de recursos pré-existentes (ex.: documentos, notícias, *logs* de aplicações etc.). Entretanto, estes modelos pré-existentes engessam o contexto de domínio, o que pode ser um problema em ambientes dinâmicos, como é o caso das discussões em *softwares* colaborativos.

O uso do contexto pode melhorar a relevância dos resultados a partir de ajustes na consulta do usuário. Estes ajustes podem ser, por exemplo, o uso da técnica de expansão de consultas. A técnica de expansão de consultas, citada na área de recuperação de informação (CARPINETO e ROMANO, 2012) e adotada neste tabalho, consiste em adicionar termos à consulta original, a fim de diminuir a ambiguação e promover maior acurácia nos resultados. Quanto mais termos, e mais representativos estes termos forem, maior a possibilidade de encontrar documentos relevantes (YATES e NETO, 1999).

A intenção é tornar a recuperação de informação sensível ao contexto das discussões (por intermédio do uso das mensagens para modelagem do contexto), oferecendo, portanto, resultados de busca contextualizados. Dey (2001) define sistemas sensíveis ao contexto como aqueles que utilizam informações de contexto para fornecer

elementos relevantes e no momento apropriado aos usuários. Segundo Bazire e Brézillon (2005), estes sistemas se adaptam a circunstâncias do ambiente, sem a interferência explícita do usuário e são úteis para caracterizar uma situação entre agentes humanos e computacionais. Por sua vez, informações de contexto, ou simplesmente contexto (DEY e ABOWD, 1999), correspondem a qualquer informação que possa ser utilizada para caracterizar a situação de uma entidade. Uma entidade é uma pessoa, lugar ou objeto que é considerado relevante para a interação entre o usuário e a aplicação, incluindo o usuário e a aplicação. Contexto, seguindo as formas descritas em (BRÉZILLON e POMEROL, 2001), e adaptadas em relação a este trabalho, representa (i) o histórico das mensagens publicadas durante um período de tempo pelos participantes do grupo, (ii) a semântica dos termos utilizados nas mensagens trocadas entre os participantes e (iii) a semântica das intenções de pesquisa.

O fato de muitos estudantes já usarem sítios de redes sociais (69,3% se enquadra na faixa etária entre 16 e 35 anos (IBOPE, 2011)), fez com que professores começassem a se familiarizar com esta tendência, para usá-la a seu favor, explorando novas possibilidades para melhorar os resultados da aprendizagem. Existem muitos exemplos de sítios de redes sociais usadas por professores e alunos como provedores de comunicação (THOMPSON, 2006) (FRANKLIN e HARMELEN, 2007) (RICHARDSON, 2009) (MORA-SOTO, 2009) (MANSUR *et* al., 2011) (DOTTA, 2011) (NASCIMENTO *et* al., 2011) (PECHI, 2011) (WANG *et* al., 2011) (BRAZ, 2011). Sítios de redes sociais geralmente fornecem recursos para o compartilhamento de conteúdo, como a publicação de documentos e *links*, e também a troca de mensagens usando programas de comunicação. Estas funcionalidades permitem o uso destes sítios como um ambiente oportuno à aprendizagem colaborativa. Em algumas plataformas de redes sociais é possível criar grupos de usuários que compartilham o mesmo interesse (por exemplo, alunos de uma classe ou um curso). Compartilhamento de conteúdo

permite fornecer o material de aprendizagem necessário para um curso, enquanto o programa de comunicação permite a troca de ideias e, portanto, tarefas colaborativas.

Estes ambientes possuem características que dificultam a captura do contexto, como ausência de conteúdo no início das discussões e mensagens escritas de maneira informal, com o uso de abreviações e linguagem coloquial, expressas com poucas palavras. Para transpor esse obstáculo, é proposta a captura do contexto a partir do enriquecimento destas mensagens em dados abertos e o uso deste contexto para melhorar a consulta do usuário, fornecendo-lhe conteúdos mais adequados a partir da *Web*. O enriquecimento do contexto é importante para acrescentar termos correlatos, visto que, em geral, a quantidade de texto presente nas mensagens em redes sociais é pequena se comparada a outros conteúdos digitais, como livros, dissertações, artigos, apostilas, páginas da *Web* etc. Portanto, o simples uso de medidas de frequência para extrair termos relevantes em uma discussão pode não ser suficiente para prover bons resultados, principalmente no início das discussões. Assim, este trabalho deve lidar com problema do arranque a frio (*cold start problem*). Tal como Chen e Sycara (1997), Smrž e Schmidt (2009), e Narayan (2009), o caráter incremental para a tarefa de extração de informação é considerado, por operar sobre um ambiente que cresce gradualmente, como os grupos de discussão em plataformas de rede social.

Para o processamento das consultas, buscaram-se algoritmos que analisassem a semelhança conceitual entre a expressão de busca e os documentos que compõem o contexto enriquecido. Isso faz com que os termos sugeridos sejam sensíveis ao contexto de domínio e relacionados à expressão de busca.

Na literatura é possível encontrar muitas propostas que visam a melhoria da recuperação da informação, com o uso de expansão de consultas sensíveis ao contexto. A abordagem apresentada nesta dissertação difere das demais na origem dos dados

(discussões em plataformas de rede social) e na forma em que o contexto é modelado (enriquecimento das discussões em dados abertos).

1.5. Metodologia de pesquisa

O primeiro passo desta pesquisa foi elaborar a revisão bibliográfica e obter o conhecimento atualizado ("estado da arte") nas áreas de recuperação de informação contextual e *Web* Semântica Social, considerando principalmente aqueles trabalhos com enfoque na educação. Foram criadas fichas de leitura para os trabalhos correlatos. A leitura crítica sobre estes trabalhos gerou uma lista de questionamentos, que fundamentam e diferenciam a presente proposta das demais. Todos os trabalhos relacionados e citados nessa pesquisa foram catalogados no *software* Mendeley[9], para que pudessem ser facilmente encontrados, tanto pela busca por palavras-chave, quanto por *tags*[10] definidas pelo autor (ex., produção-unirio, linked-data, colaboração, avaliação, recuperação-informação etc.). Cada documento foi identificado por uma ou mais *tags*.

A partir do desenvolvimento da pesquisa foram elaborados e executados dois estudos de caso. Para cada estudo foi necessário desenvolver um protótipo de sistema de informação, responsável por coletar os dados de uso, oriundos da interação aluno-computador, em um ambiente de aprendizagem.

Tanto o primeiro estudo de caso (PRATES *et* al., 2011) (PRATES *et* al., 2012) quanto o segundo foram realizados em disciplinas de graduação da UNIRIO e contaram com a participação de estudantes, além do professor e pesquisador, que atuaram como mediadores da dinâmica.

[9] http://www.mendeley.com/
[10] *Tag* é um rótulo relevante que descreve uma informação, como os artigo no âmbito deste trabalho.

O primeiro estudo de caso foi elaborado para investigar a expansão de consultas a partir de documentos representativos do contexto e de discussões em grupos de redes sociais. Alterações, decorrentes das observações realizadas no primeiro estudo de caso (Capítulo 4), foram realizadas tanto na proposta (Capítulo 5), quanto no processo de avaliação e culminaram na execução do segundo estudo de caso (Capítulo 6). O segundo estudo de caso aprofundou a investigação sobre o contexto a partir de discussões, enriquecendo os termos extraídos com dados abertos.

O método estudo de caso foi adotado, pois esta é uma pesquisa empírica [11] realizada em um contexto real, e, portanto, sem controle total sobre os acontecimentos e variáveis (YIN, 2005). Variável é um rótulo atribuído a um fenômeno que deve ser controlado objetivamente. Este fenômeno é observável, suscetível a alterações e mensurável (quando possível) ao longo de uma pesquisa. Três tipos de variáveis foram utilizadas nessa pesquisa: independentes, dependentes e intervenientes (WAZLAWICK, 2009). A variável independente (ou experimental) é aquela que sofre manipulação controlada e proposital por parte do pesquisador, e, com isso, pode influenciar, determinar ou afetar outra variável (ou outras variáveis). As variáveis independentes consideradas foram: (i) a quantidade de documentos retornados pela consulta, (ii) a quantidade de *clusters* e (iii) a forma de extração dos termos (geral ou cluster), bem como o número máximo de termos a serem usados na expansão da consulta (primeiro estudo de caso) e (iv) o número de grupos e termos associados a estes grupos, pois cada grupo possui uma forma de extração diferenciada (segundo estudo de caso). A variável dependente é aquela que sofre influência de outras variáveis (causa), como a manipulação da variável independente, que deve ser observada ou medida (efeito). A variável dependente identificada foi a avaliação de relevância (utilidade) aferida pelos

[11] Segundo Kerlinger (1980), empírico significa "guiado pela evidência obtida em pesquisa científica sistemática e controlada".

alunos para os resultados obtidos no protótipo para suas buscas. Essa variável (relevância) é composta por um conjunto de valores discretos e categóricos (0=irrelevante[12], 1=pouco relevante[13], 2=parcialmente relevante[14], 3=relevante[15] e 4=totalmente relevante[16]). Ainda, para o segundo estudo de caso, considerou-se o grupo de termos selecionado pelo usuário para a expansão da consulta (associação entre termo selecionado e grupo a que pertence este termo). Já a variável interveniente[17], é aquela que pode interferir na relação de causalidade entre variáveis dependentes e independentes, mas não pode ser manipulada, medida e suprimida pelo pesquisador. As variáveis intervenientes consideradas foram o perfil e conhecimento sobre o assunto abordado pelos alunos selecionados para cada estudo de caso.

As avaliações coletadas no protótipo são recursos que representam julgamento humano de relevância para cada documento *Web* retornado. As métricas, utilizadas para comparar os resultados da busca original e expandida, foram: (i) precisão total dos *x* primeiros resultados (CHIGNELL *et al*., 1999), (ii) comprimento da busca (COOPER, 1968) e (iii) correlação de *ranking* (SU *et al*., 1998). A precisão total dos *x* primeiros resultados (*first x full precision*) mede a quantidade total de informações relevantes nos *x* primeiros documentos. O comprimento da busca (*search length*) mede o número de documentos não relevantes que um usuário deve examinar antes de encontrar uma quantidade *x* de documentos relevantes. Por fim, correlação de *ranking* (*rank correlation*) mede a correlação entre a classificação do sistema e o julgamento do usuário para os resultados da busca. Espera-se que ocorra dependência direta entre a

[12] Documento absolutamente irrelevante e fora do contexto do tema pesquisado.
[13] Pouco relevante e que não supra a necessidade de informação, mas com alguma relação ao tema pesquisado.
[14] Supra parcialmente a necessidade de informação ou contenha um link para outro documento de classificação maior.
[15] Documento suficiente para suprir a necessidade de informação.
[16] Documento que supra a necessidade de informação.
[17] Conforme Lakatos e Marconi (2001), a variável interveniente é aquela que, "numa sequência causal, coloca-se entre a variável independente (I) e a variável dependente (D), tendo a função de ampliar, anular ou diminuir a influência de I sobre D".

variável independente “grupo dos termos expandidos” e a variável dependente de avaliação dos resultados, com superioridade da consulta expandida em relação à consulta original.

Ao final de cada estudo, um questionário foi disponibilizado aos participantes em meio eletrônico e buscou confirmações qualitativas sobre os resultados quantitativos das avaliações. Perguntou-se, por exemplo, sobre “qual grupo de termos trouxe melhores resultados?” e também se a “recomendação de termos melhorou com o tempo”. Essas respostas foram trianguladas com os resultados das avaliações coletadas no protótipo (análise de *log*).

1.6. Objetivos da dissertação

O objetivo desta dissertação é prover a recuperação de conteúdos na *Web* que sejam relevantes durante dinâmicas de aprendizagem colaborativa, usando informação contextual do domínio, obtida a partir do enriquecimento das mensagens em redes sociais. A proposta enfoca situações onde não exista ou não se deseje obter esforço de especialistas para modelagem do contexto de domínio. No âmbito educacional, vários trabalhos mostram a importância de orientar os alunos na aquisição de recursos de informação e a dificuldade em recomendar recursos na *Web* (KUIPER *et* al., 2005) (RENSING *et* al., 2010). Portanto, seria interessante fornecer funcionalidades de pesquisa na *Web* (ZIMMER, 2010) em ambientes de aprendizagem. Com isso, o objetivo específico desta dissertação é auxiliar os alunos durante dinâmicas de aprendizagem colaborativa baseada em discussão, a partir de grupos de rede social, com a pretensão de prover-lhes documentos da Web que supram suas necessidades de informação. Dinâmicas de aprendizagem colaborativa baseadas em discussão tornam o problema ainda mais evidente, visto que, nesse tipo de abordagem, a colaboração entre

os alunos, coordenadas pelo professor, é essencial ao aprendizado e a recuperação precisa e contextualizada de documentos na Web e poderia auxiliá-los durante a dinâmica. Para o cumprimento do objetivo principal, as seguintes tarefas foram traçadas e desenvolvidas:

- Utilizar técnicas de recuperação de informação e processamento de linguagem natural para tratamento de mensagens de redes sociais;
- Tratar a falta de informação no início das discussões;
- Enriquecer os termos provenientes de discussões, a partir do uso de dados abertos para a identificação de conceitos e expansão das instâncias desses conceitos pelas relações de similaridade semântica entre eles;
- Criar regras de extração e sugestão de termos para a expansão de consultas sobre o contexto enriquecido;
- Automatizar o processo de indexação e cálculo de frequência para os termos;
- Desenvolver um protótipo de *software* que cumpra os requisitos da proposta de solução e permeie a avaliação do processo como um todo;

1.7. Organização da dissertação

Com o propósito de orientar o leitor sobre a distribuição dos assuntos, os capítulos dessa dissertação foram organizados da seguinte forma (Figura 2):

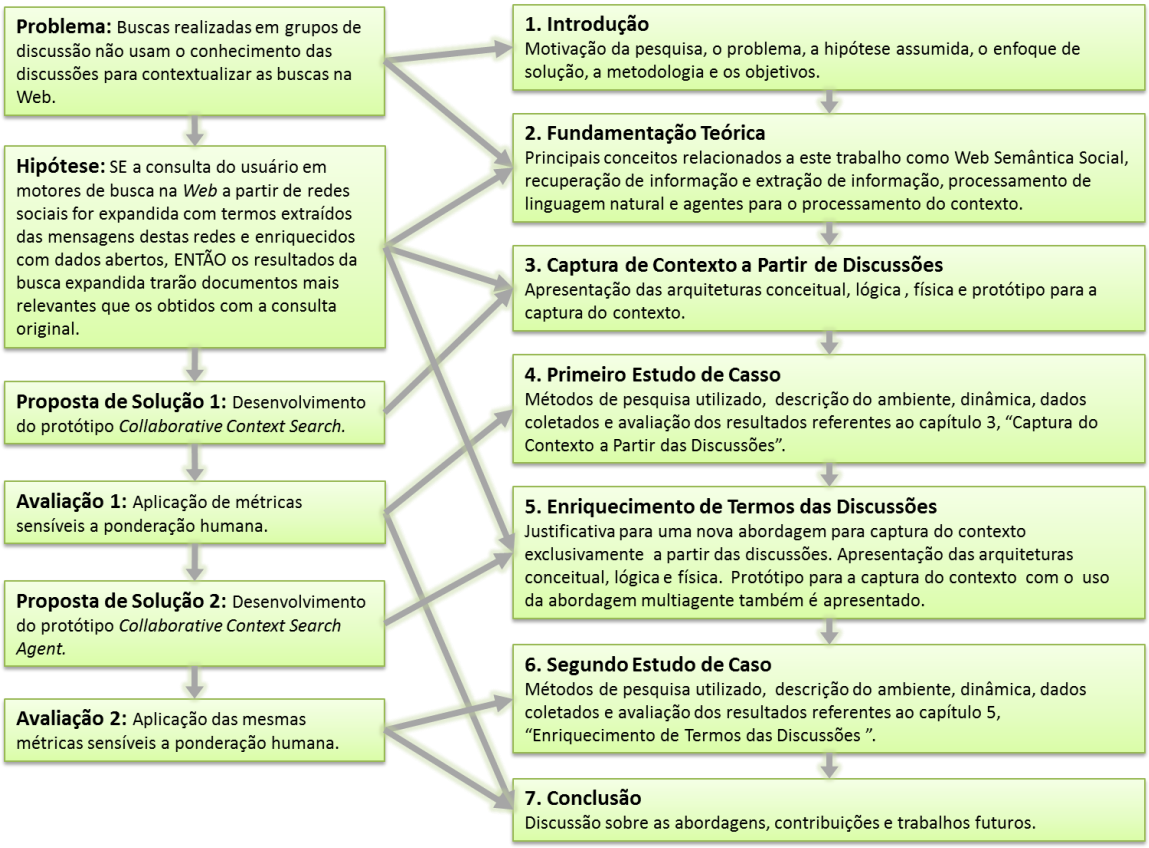


Figura 2 – Organização da Dissertação

Capítulo 2 – Fundamentação Teórica

Neste capítulo serão apresentadas algumas definições basilares para a compreensão deste trabalho.

2.1. *Web* Social: *Wikis* e Redes Sociais Online

A *Web* Social é o rótulo para o fenômeno relacionado à mudança de paradigma sobre a autoria dos conteúdos na *Web*. O conteúdo passa a ser gerado e gerido por pessoas a partir de aplicações específicas baseadas na *Web*, como as *wikis* e as redes sociais, que estão relacionadas a este trabalho. De maneira geral, estas aplicações possibilitam o compartilhamento de informações e promovem a interação social entre estas pessoas e seus conteúdos.

As redes sociais online permitem conectar pessoas de forma descentralizada na *Web*. As pessoas são descritas por um perfil, o qual contempla informações pessoais, hábitos e preferências, disponíveis e utilizadas como referência para a criação de novos laços sociais. Plataformas de redes sociais oferecem meios de interação interpessoal e de comunicação para partilha de informação e conhecimento de interesse comum, a partir de critérios de afinidade (FRANKLIN e HARMELEN, 2007). Existem muitos exemplos de aplicações *Web* sociais (ou sítios de redes sociais), cada qual criada para atender propósitos diferentes, como o *Twitter* para microblogs, o *Youtube* para vídeos, o *Flickr* para fotos, o *Linkedin* para contatos profissionais e o *MySpace*, *Facebook* e *Orkut* para amizades em geral.

No caso das *wikis*, como a enciclopédia multidomínio *Wikipédia*, os usuários não precisam saber códigos complexos para gerar conteúdo. Cada um interage, por intermédio de interfaces fáceis de usar e interativas, a fim de contribuir com a criação coletiva de tópicos, que somados resultam em uma base de conhecimento compartilhada, também chamada de inteligência coletiva (LÉVY, 1999) ou inteligência acumulada (GRUBER, 2008). A qualidade do conteúdo melhora à medida que mais pessoas colaboram, ou seja, trata-se de um grande esforço para a coprodução e difusão do conhecimento (MIKROYANNIDIS, 2007). Todo o conhecimento é criado e mantido pela própria comunidade, sem nenhum controle rígido, de acordo com os moldes da *Web* Social (OREILLY, 2007). A *Wikipédia* possui mais de 718 mil artigos em português e 3,9 milhões em inglês. Assim como a *Web* sintática, o valor semântico[18] de todo o seu conteúdo depende da leitura e interpretação humana da linguagem, ou seja, não é endereçado a máquinas, mas sim a pessoas.

2.2. *Web* Semântica

A *Web* abrange um amplo domínio de conhecimento e envolve questões terminológicas conflitantes. A representação inequívoca do conhecimento, com seus conceitos, atributos e relacionamentos pode organizar um domínio de conhecimento e reduzir a dupla interpretação para os termos usados nesse domínio (ANTONIOU E HARMELEN, 2004). A *Web* Semântica, vislumbrada por Berners-Lee, Hendler, e Lassila, surge com a pretenção de estruturar e dar significado para as coisas, com o objetivo de possibilitar a extração e processamento de informações por máquinas (BERNERS-LEE, *et al*., 2001).

[18] Semântica é o ramo da linguística que explora o significado das coisas

A *Web* Semântica pode ser entendida como um novo conjunto de práticas e tecnologias que são somadas às da *Web* atual, para que a visão de Berners-Lee, Hendler e Lassila exista. Essas tecnologias possibilitam a ligação de dados (e não somente documentos/páginas) na *Web* e a associação desses dados aos modelos conceituais que os descrevem. Esses modelos, chamados de Ontologias (seção 2.2.1), são descritos em linguagem formal e dão suporte a tarefas de raciocínio computacional automatizado e interoperabilidade semântica.

Dentre as muitas aplicabilidades, pode ser destacada (i) a melhoria na precisão dos resultados às buscas na *Web* (PINHEIRO, 2004), (ii) recuperação de fontes de informação em repositórios de metadados (SANTOS, 2011), (iii) recuperação de mensagens de discussão em comunidades de prática (VARELLA, 2007), (iv) o auxílio ao processamento de linguagem natural (MOTTA, 2009), (v) o suporte à gestão do conhecimento (GATTI, 2009) (GATTI *et* al., 2010) e (vi) o suporte aos sistemas educacionais (SIQUEIRA, 2005) (MATTOS, 2006) (ISOTANI *et* al., 2009) (OLIVEIRA, 2009) (BITTENCOURT, 2009) (BITTENCOURT e COSTA, 2011) (HERLI, 2011) (GLUZ e VICCARI, 2011) (BRAZ *et* al., 2011). Nos próximos tópicos, abordar-se-ão mais detalhes sobre as tecnologias da *Web* Semântica empregadas neste trabalho.

2.2.1. Ontologias

Ontologias, sob a perspectiva computacional, são modelos de referência concebidos para representar conceitos (conhecimento sobre o mundo) de forma padronizada e consistente (MIZOGUCHI e KITAMURA, 2004), usando vocabulário comum e consensual (GRUBER, 1993). Tradicionalmente, a modelagem do domínio de conhecimento é feita por especialistas de domínio e engenheiros de conhecimento que usam métodos, técnicas e linguagens da engenharia de ontologias para a definição de

conceitos e relações entre conceitos do modelo (FERNANDEZ-LOPEZ e CORCHO, 2004).

O modelo deve aderir a uma linguagem formal que torne possível a correta interpretação dos conceitos, seus atributos, propriedades e restrições. Segundo Breitman (2005), "ontologias são modelos conceituais que capturam e explicitam o vocabulário utilizado nas aplicações semânticas. Servem como base para garantir uma comunicação livre de ambiguidades e será a língua franca da *Web* Semântica".

2.2.2. Linguagens para Construção de Ontologias

As ontologias são descritas por linguagens padronizadas de metadados. Para se construir uma ontologia é necessário decidir qual linguagem melhor adequa-se a necessidade da modelagem, tendo em vista que, quanto maior a expressividade da linguagem, maior a riqueza para representação dos conceitos, porém, menor a decidibilidade e tratabilidade computacional.

Nesse âmbito, o XML [19] (*Extensible Markup Language*) oferece uma metalinguagem de marcação para a criação estruturada de documentos, por intermédio de *tags* (marcadores). A criação de *tags* é livre e deve obedecer a um conjunto de regras sintáticas. Entretanto, a liberdade dos usuários para a criação de *tags* no XML pode gerar ambiguidades para a definição dos conceitos. Para eliminar esta ambiguidade, vocabulários foram definidos e padronizados pela W3C (*World Wide Web Consortium*), como, por exemplo, XML Schema[20], RDF (*Resource Description Framework)*[21], RDFS (*Resource Description Framework Schema*)[22] e OWL (*Web Ontology Language*)[23].

[19] http://www.w3.org/XML/
[20] http://www.w3.org/XML/Schema
[21] http://www.w3.org/RDF/
[22] http://www.w3.org/TR/rdf-schema/
[23] http://www.w3.org/TR/owl-features/

O RDF é uma linguagem que oferece um modelo simplificado para descrever recursos na *Web* (W3C, 2004). A padronização da descrição destes recursos permite a sua leitura e interpretação por computadores. Usa sintaxe XML, porém outras representações sintáticas são possíveis, como, por exemplo, N-Triples[24], N3/Turtle[25] e JSON[26]. Esse modelo é formado por declarações de três componentes (triplas). Cada tripla é composta por um sujeito, um predicado e um objeto (*Subject-Predicate-Object*), nesta ordem, e define uma afirmação sobre algo. O sujeito identifica a entidade ou recurso em questão, que pode ser qualquer coisa (*thing*) que possua um URI[27]. O predicado é um atributo ou propriedade que descreve relações entre recursos e também é identificado por um URI. O objeto é o valor que é atribuído a um sujeito e pode ser um valor literal primitivo (cadeia de caracteres, número, data) ou outro recurso.

Enquanto o RDF fornece uma maneira de expressar declarações simples sobre recursos, outras linguagens como o RDFS e o OWL proveem maior representatividade e permitem a criação de esquemas de classes, hierarquias, propriedades, relações, restrições e padronizações terminológicas. O RDFS permite a modelagem de ontologias simples, enquanto o OWL possui três versões: OWL Lite (mapeamento de tesauros), OWL-DL (lógica de descrição) e OWL *Full* (grande expressividade lógica, porém a eficiência computacional não é garantida).

2.2.3. SPARQL

SPARQL (*SPARQL Protocol and RDF Query Language*) é a especificação de uma linguagem de consulta para recuperação de dados em RDF, recomendada pelo consórcio W3C (PRUD'HOMMEAUX e SEABORNE, 2008) e que lembra em diversos

[24] http://www.w3.org/2001/sw/RDFCore/ntriples/
[25] http://www.w3.org/TeamSubmission/turtle/
[26] http://www.json.org/
[27] Identificador Uniforme de Recursos (Uniform Resource Identifier, em inglês) é um descritor usado para identificar recursos na internet de maneira singular.

aspectos a linguagem SQL tradicional (KEßLER, 2010). RDF e SPARQL são tecnologias essenciais da *Web* Semântica para a publicação e recuperação de dados ligados (BIZER *et* al., 2009).

2.2.4. Dados ligados abertos

Dados ligados abertos (*linked open data*) refere-se à criação de conexões entre dados na *Web* e à disponibilização destes dados de maneira não proprietária. Inicialmente, a *Web* foi concebida para utilizar o par URI/HTTP para identificar documentos, agora, passa a servir também para descrever ligações ricas entre os diferentes recursos (coisas, pessoas, lugares, entre outros), em formato adequado, e disponíveis, por meio de URIs (BIZER, *et* al., 2007). Para a semântica entre essas ligações é utilizada a linguagem RDF. Definir explicitamente o significado dos relacionamentos favorece a criação de uma rede semântica de dados. A nuvem formada a partir deste elo entre os elementos faz com que as máquinas possam desempenhar um papel ativo na *Web* (BIZER *et* al., 2009). Esforços, como o *LinkingOpenData*[28] (Figura 3), nascem na tentativa de viabilizar a publicação de dados RDF e a interligação entre diferentes fontes de dados abertos, e, além disso, possuem o objetivo de tornar este trabalho disponível gratuitamente a todos (BIZER *et* al., 2009).

O pesquisador e evangelista Berners-Lee, no evento TED[29], em fevereiro de 2009 reforça essa ideia ao afirmar que os dados devem ser disponibilizados em seu estado natural, sem tratamento, em formato compreensível por máquinas. Para tanto, é necessário que se evitem os silos de dados, disponibilizando-os em um formato padrão.

[28] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[29] http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

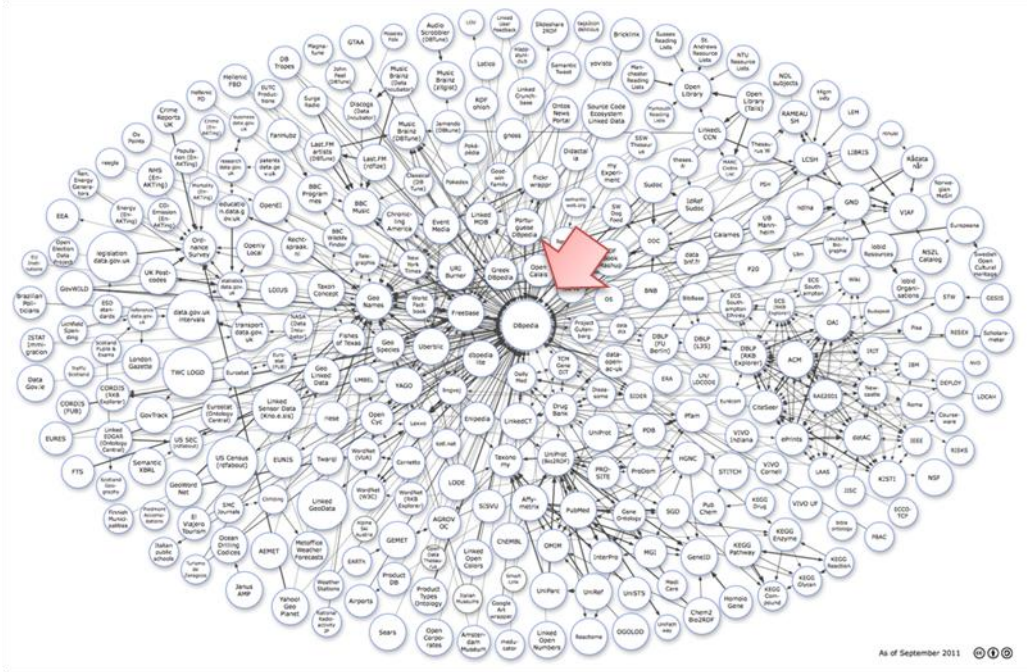


Figura 3 – *DBpedia* (ao centro) no Contexto "*Linking Open Data*" (LOD)

2.2.5. DBpedia

O projeto *DBpedia*[30] é um esforço da comunidade para extrair dados estruturados da *Wikipédia* e organizá-los na forma de dados interligados e abertos. O funcionamento da *DBpedia* consiste na extração (*dump*) dos *infoboxes*[31] (ou "caixas de informação") que apresentam dados estruturados encontrados na *Wikipédia* (Figura 4), e seu posterior mapeamento em ontologias. A filosofia de ligação de dados utiliza o modelo RDF e, com isso, permite a criação de consultas ricas, baseadas em SPARQL (AUER, 2007).

[30] http://wiki.*DBpedia*.org/About
[31] http://wiki.dbpedia.org/Datasets

Município de Criciúma

"Capital do Carvão"

Fundação	6 de janeiro de 1880 (132 anos)
Gentílico	*criciumense*
Prefeito(a)	Clésio Salvaro (PSDB) (2009–2012)
Localização	
	28° 40' 40" S 49° 22' 12" O
Unidade federativa	Santa Catarina
Municípios limítrofes	Siderópolis, Cocal do Sul, Morro da Fumaça, Maracajá, Araranguá, Nova Veneza, Forquilhinha, Içara
Distância até a capital	191 km
Características geográficas	
Área	235,628 km² *(BR: 38939)*[2]
População	193 988 hab. *(SC: 5º)* – *Estimativa IBGE/2011*[3]
Densidade	823,28 hab./km²
Altitude	46 m
Clima	subtropical Cfa
Fuso horário	UTC−3
Indicadores	
IDH	0,822 *elevado PNUD/2000*[4]
PIB	R$ 2 791 692,467 mil *IBGE/2008*[5]
PIB per capita	R$ 14 927,40 *IBGE/2008*[5]

Figura 4 – Exemplo de um *infobox* da *Wikipédia*

A Figura 5 apresenta uma instância da *DBpedia* denominada "Fluxo de Informação", obtida a partir do processamento dos dados da *Wikipédia*.

About: Fluxo de informação

An Entity of Type : Thing, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

O conceito de Fluxo de Informação é utilizado por três campos diferentes de conhecimento: a Semiótica, que considera a influência dos fluxos na construção do discurso; a Teoria da Informação, fortemente influenciada por modelos matemáticos e de informática; e a Teoria da Comunicação, que identifica tais fluxos com a organização geopolítica e geocultural do mundo. fluxos são setas,

Property	Value
dbpedia-owl:abstract	• In discourse-based grammatical theory, information flow is any tracking of referential information by speakers. Information may be new, just introduced into the conversation; given, already active in the speakers' consciousness; or old, no longer active. The various types of activation, and how these are defined, are model-dependent. Information flow affects grammatical structures such as word order. active, passive, or middle voice. choice of deixis, such as articles; "medial" deictics such as Spanish ese and Japanese sore are generally determined by the familiarity of a referent rather than by physical distance. overtness of information, such as whether an argument of a verb is indicated by a lexical noun phrase, a pronoun, or not mentioned at all. • O conceito de Fluxo de Informação é utilizado por três campos diferentes de conhecimento: a Semiótica, que considera a influência dos fluxos na construção do discurso; a Teoria da Informação, fortemente influenciada por modelos matemáticos e de informática; e a Teoria da Comunicação, que identifica tais fluxos com a organização geopolítica e geocultural do mundo. fluxos são setas,
dbpprop:auto	• yes
dbpprop:date	• December 2009
dbpprop:wikiPageUsesTemplate	• dbpedia:Template:Unreferenced
dcterms:subject	• category:Discourse_analysis • category:Information_science
rdfs:comment	• O conceito de Fluxo de Informação é utilizado por três campos diferentes de conhecimento: a Semiótica, que considera a influência dos fluxos na construção do discurso; a Teoria da Informação, fortemente influenciada por modelos matemáticos e de informática; e a Teoria da Comunicação, que identifica tais fluxos com a organização geopolítica e geocultural do mundo. fluxos são setas, • In discourse-based grammatical theory, information flow is any tracking of referential information by speakers. Information may be new, just introduced into the conversation; given, already active in the speakers' consciousness; or old, no longer active. The various types of activation, and how these are defined, are model-dependent. Information flow affects grammatical structures such as word order. active, passive, or middle voice.
rdfs:label	• Information flow • Fluxo de informação
owl:sameAs	• fbase:Fluxo de informação
foaf:page	• http://en.wikipedia.org/wiki/Information_flow
is dbpedia-owl:wikiPageDisambiguates of	• dbpedia:Information_flow_(disambiguation)
is owl:sameAs of	• http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Information_flow
is foaf:primaryTopic of	• http://en.wikipedia.org/wiki/Information_flow

Browse using: OpenLink Data Explorer | Zitgist Data Viewer | Marbles | DISCO | Tabulator Raw Data in: CSV | RDF (N-Triples N3/Turtle JSON XML) | OData (Atom JSON)| Microdata (JSON HTML) | JSON-LD

About

This content was extracted from Wikipedia and is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License

Figura 5 – Instância "Fluxo de Informação" na *DBpedia*

2.3. Sistemas de Recuperação da Informação

Segundo Manning *et* al. (2008), sistemas de recuperação de informação (SRI) são programas de computador responsáveis pelas atividades de indexação, busca e classificação de grandes coleções de documentos em formato digital. Estas coleções, também chamadas de *corpus* ou *corpora*, são compostas por conteúdo textual descrito em linguagem natural, ou seja, em formato não estruturado.

O objetivo principal de um SRI é promover a recuperação da informação, ou seja, encontrar o documento (ou um subconjunto de documentos) relevante no *corpus*. Documentos relevantes são aqueles que atendem à necessidade de informação definida pelo usuário. Em geral, a necessidade de informação é expressa em uma linguagem de consulta, composta por palavras-chave e operadores lógicos. Outros critérios de consulta avançados, como idioma da pesquisa, também podem ser definidos.

Para auxiliar o processo de busca de informação, o usuário tem a sua disposição os chamados "motores de busca" ("*search engines*" em inglês) ou, popularmente, "buscadores". Segundo dados de (ALEXA, 2012), o mecaniso de busca do Google é o sítio com maior audiência na *Web*, seguido pelo sítio da plataforma de rede social *Facebook*. Este fato nos faz refletir a importância dos buscadores no cotidiano das pessoas.

Encontrar a informação relevante é, entretanto, uma tarefa difícil. Palavras podem ter muitos significados de acordo com o contexto em que são empregadas. Buscas por palavras-chave tendem a ser imprecisas, pois não são capazes de capturar a semântica, nem o contexto que se deseja.

2.3.1. Abordagens para Recuperação de Informação

Os sistemas de recuperação de informação usam modelos de representação para a expressão de busca do usuário e para os documentos que são indexados. Dentre os modelos encontrados na literatura, ressaltam-se o booleano, o espaço vetorial e o de análise semântica latente.

2.3.1.1. Modelo Booleano

O modelo booleano considera a teoria de conjuntos e a álgebra booleana como estratégia de recuperação de documentos. Os termos da expressão de busca devem ser ligados aos operadores lógicos AND, OR ou NOT. Dessa maneira, a relação de pertinência para os documentos recuperados é binária. Os documentos recuperados são apenas aqueles que satisfazem a expressão lógica da consulta. Não há possibilidade de casamento parcial entre termos e documentos. Também não há possibilidade de estabelecer critérios de ordenação para o conjunto de resposta. Minimamente, o usuário deve restringir ao máximo seu conjunto resposta, por meio de uma consulta bem formulada (MANNING *et* al., 2008).

2.3.1.2. Modelo de espaço vetorial

O modelo de espaço vetorial (SALTON, 1989) surge para resolver certas limitações da abordagem booleana, como a necessidade de ordenação do resultado. A ordenação é fundamental nos casos em que o conjunto de resultado é muito grande, e pode marcar o sucesso da busca.

Para fins de indexação, os documentos são representados como vetores multidimensionais. Para cada termo de cada documento é realizado o cálculo dos pesos, como, por exemplo, a sua frequência no texto. Nesse espaço multidimensional, se um termo não ocorre no documento, o valor do peso será zero. O conteúdo textual é

representado como um “saco de palavras” (ou *bag-of-words*, em inglês), portanto não há distinção entre a posição em que as palavras ocorrem no texto.

O sucesso da recuperação está diretamente relacionado à função de similaridade entre os termos da consulta e o índice da coleção. Essa função de similaridade é fundamentada nos princípios da álgebra linear e permite o casamento parcial entre os vetores de consulta e documentos. Isso define a ordem em que os documentos são exibidos ao usuário (MANNING *et* al., 2008).

2.3.1.3. Análise Semântica Latente

Análise semântica latente (ASL) (DEERWESTER, DUMAIS e HARSHMAN, 1990), também chamada de Indexação de Semântica Latente (em inglês, *Latent Semantic Analysis* e *Latent Semantic Indexing*, respectivamente) é um método moderno para recuperação de informações (MANNING *et* al., 2008). A ASL parte do pressuposto de que a necessidade de informação tem maior relação com os conceitos - ou ideias - do que com os termos do índice somente. Existem diversas maneiras de expressar, em palavras, um mesmo conceito no mundo. A análise entre os relacionamentos destas palavras define o conceito (FOLTZ, 1996).

A representação computacional destes conceitos deve lidar com problemas de significação das palavras (campo semântico), como a sinonímia [32] e polissemia [33] (DEERWESTER *et al*., 1990). Por exemplo, no contexto do curso de sistemas de informação, a consulta “banco de dados” retornará documentos associados ao conceito “conjunto de registros gerenciados por um SGBD”. Portanto, documentos com a expressão “base de dados”, “base relacional” ou simplesmente “SGBD” também poderão ser retornados, mesmo se as palavras “banco” ou “dados” não existirem nestes

[32] Qualidade das palavras que tem exatamente o mesmo sentido que outra ou quase idêntico.
[33] Qualidade das palavras que variam de sentido

documentos. Por outro ponto de vista, mesmo se os termos “banco” e “dados” façam parte de um documento, este não será considerado se a interpretação destes termos estiver relacionada a outros conceitos (como “dados” usado em jogos de tabuleiro ou “banco” para doação de sangue).

O espaço conceito ou espaço semântico (DEERWESTER *et* al., 1990) procura estabelecer relações de similaridades conceituais latentes (ocultas). Para a criação desse espaço, primeiro deve-se criar uma matriz que defina a coocorrência entre termos e documentos. Em seguida, essa matriz deve passar por um processo de decomposição em valores singulares, para reduzir a dimensionalidade desses vetores (BERRY *et* al., 1999). Tanto os termos que expressam a consulta quanto os documentos devem ser convertidos para o espaço conceitual e representados como vetores nesse espaço. A recuperação de documentos segue critérios de similaridade, definida pelo ângulo cosseno entre os vetores da consulta e dos documentos (BERRY *et* al., 1999).

Este trabalho explora as potencialidades do método de análise de semântica latente, que busca melhorar a recuperação de documentos, com base nos termos da consulta informados pelo usuário.

2.3.2. Cálculo de pesos

O cálculo dos pesos define a importância (valor numérico) de todos os termos, que ocorrem em consultas ou documentos, no espaço vetorial. *Term Frequency (TF)* e *Term Frequency – Inverse Document Frequency (TF-IDF)* são maneiras de se calcular os pesos, que compartilham a ideia de contar a frequência dos termos no texto.

Utiliza-se o *TF* para evidenciar os termos mais relevantes e que melhor representem um documento, a partir do número de ocorrências de termos neste documento. A aplicação do logaritmo na fórmula de frequência de termos (*sublinear tf scaling*) busca diminuir a influência desta frequência, conforme descrito em

(MANNING *et* al., 2008). A seguinte equação ilustra a aplicação do logaritmo (*log*) no cálculo da frequência de termos, onde: $wf_{t,d}$ representa o peso de um termo (t) em um documento (d), e $tf_{t,d}$ representa a frequencia de um termo (t) em um documento (d).

$$\mathrm{wf}_{t,d} = \begin{cases} 1 + \log \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Utiliza-se a métrica *TF–IDF* para diferenciar um termo dentro de uma coleção, ou seja, demonstrar a importância (ou originalidade) de um termo frente o *corpus*. Esta medida aumenta se o termo ocorre em um documento e diminui se ocorre no *corpus* com grande frequência, denotando assim sua originalidade.

2.3.3. Expansão de consultas

Em geral, os usuários devem ajudar o buscador refinando sua expressão de busca, acrescentando ou removendo palavras-chave. Novas consultas deverão ser feitas até que sua necessidade de informação seja atendida. Entretanto, essa não é uma tarefa trivial, e pode levar o usuário ao estado de frustração (ALLAN *et* al., 2005). Para esse problema, utiliza-se a técnica de expansão de consultas, que consiste em adicionar termos relacionados à consulta original feita pelo usuário, com a proposta de melhorar a eficiência dos sistemas de recuperação de informação (RIJSBERGEN *et* al., 1998).

A motivação para o uso desta técnica é tratar o problema do descasamento de palavras (*word mismatching problem*, em inglês) entre os termos da consulta do usuário e os termos do índice de documentos (XU e CROFT, 1996). Isso ocorre, pois os termos são sintaticamente indexados, ou seja, não existe relação de significação entre as palavras. Assim, somente os documentos que possuírem o(s) termo(s) exato(s) da consulta serão retornados pelo motor de busca. Nesse caso, utiliza-se a disjunção de sinônimos (por exemplo: Carro OR Automóvel OR Veículo OR Automotor). Sinônimos nesse caso são todos os termos que representem o mesmo conceito no mundo e não,

necessariamente, os sinônimos da língua portuguesa. Por exemplo, o termo “cobertura” pode estar associado ao conceito de “recuperação de informação” ou “teste de *software*”. Nesse caso, os possíveis sinônimos da língua portuguesa, “capelo”, “capuz”, “tapume”, “capa”, “envolver”, “proteger”, “abrigar”, “resguardar” etc., não se relacionam com o conceito anterior.

A expansão de consulta também pode ser usada para restringir os resultados, pela adição de termos que especializem o tema pesquisado. Como os motores de busca assumem a definição mais popular para um termo (problema), a adição de termos pode melhorar problemas de ambiguidade. Nesse caso, usa-se a conjunção de termos (por exemplo: Cobertura AND Recuperação AND Informação).

Os termos candidatos à expansão podem ser oriundos de estruturas de representação do conhecimento (taxonomias, tesauros ou ontologias) (VOORHEES, 1994) ou então extraídos de conteúdos textuais por meio de medidas estatísticas de coocorrência dos termos (PEAT e WILLETT, 1991).

A dinâmica para a realização das consultas pode ocorrer de três maneiras: expansão manual, interativa ou automática. Na expansão automática, os termos com maior peso são adicionados à consulta sem a intervenção do usuário. Já na expansão manual (sem apoio do sistema de busca) e na interativa (com apoio do sistema de busca), os termos são sugeridos ao usuário, que deve escolher aqueles que melhor atendam sua necessidade de informação (RIJSBERGEN *et* al., 1998) (KACPRZYK *et* al., 2008). Nessa dissertação, adotou-se a expansão de consultas automática para o primeiro estudo de caso e interativa para o segundo.

2.3.4. Avaliação clássica em sistemas de recuperação de informação

Precisão e cobertura são fórmulas muito utilizadas na avaliação de sistemas de recuperação de informação, que objetivam medir a eficácia da recuperação de um

conjunto de documentos. Três requisitos são necessários para o cálculo da avaliação: (i) uma coleção de documentos, (ii) um conjunto de necessidades de informação expressas por uma consulta e (iii) um conjunto de julgamentos de relevância, para cada par consulta-documento (MANNING *et* al., 2008).

Precisão é a fração de documentos recuperados que são relevantes, enquanto cobertura a fração de documentos relevantes que são recuperados (MANNING *et* al., 2008). Neste trabalho, como não é possível estimar quais documentos são relevantes, visto que a fonte de informação em questão é a *Web*, adotaram-se outras propostas para a avaliação desta pesquisa, discutidas na seção 1.5.

2.4. Extração de Informação

A técnica de extração de informação tem por objetivo processar informações não estruturadas em busca de elementos e relações entre elementos que possam ser representados de maneira estruturada (SARAWAGI, 2008). Sua finalidade é agilizar o acesso a estas informações, visto que essa busca pode ser muito complexa e demorada se realizada por humanos. Dentre as tarefas típicas para extração de informação, relacionam-se a este trabalho o reconhecimento de entidades e a relação entre entidades (MOENS, 2006). O reconhecimento de entidades visa a extração e classificação de qualquer coisa (*thing*) do mundo real, concreta ou abstrata, mencionada no texto. Já a relação entre entidades busca por elos semânticos entre as entidades. Neste trabalho, o reconhecimento de entidades e a relação entre entidades fazem uso de processamento de linguagem natural e a ontologia de uma enciclopédia colaborativa.

2.5. Processamento de Linguagem Natural

O processamento de linguagem natural (PLN) tem por objetivo automatizar a geração e a compreensão das linguagens naturais (escrita ou falada), ou seja, transformar a linguagem humana em modelos computacionais e vice versa. A geração de modelos computacionais é importante para que haja tratabilidade na execução de programas de computador e, por conseguinte, viabilize a descoberta de conhecimento em *corpus* textual (MANNING e SCHÜTZE, 1999).

Um *corpus* textual é composto por termos, que, por sua vez, são expressos por significantes, ou seja, palavras ou combinações de palavras, que variam de acordo com a língua. Um mesmo termo pode representar um ou mais objetos do mundo, de acordo com o contexto utilizado. A Figura 6 ilustra a ambiguidade que o processamento de linguagem natural deve tratar ao estabelecer relação entre termos e conceitos.

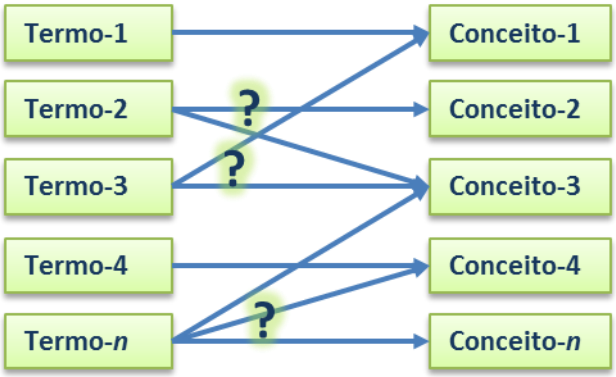


Figura 6 – Mapeamento entre termos e conceitos

A reunião de termos de um mesmo domínio, que reflita o senso comum de uma comunidade, forma um sistema terminológico conceitual dotado de significado semântico e pragmático (o conceito é o significado) (KRIEGER e FINATTO, 2001).

As tarefas de processamento de texto empregadas neste trabalho serão apresentadas a seguir. Mais informações podem ser obtidas em (MANING e SCHÜTZE, 1999).

2.5.1. Tokenização

A tokenização (ou *tokenization*) é um procedimento aplicado ao texto, que objetiva dividi-lo em unidades lexicais mínimas, intituladas *tokens* (BARRETO *et* al., 2006). Cada palavra, em geral, é um *token*. As palavras são separadas por um caracter especial de espaçamento, usado como delimitador para a geração dos *tokens*. Entretanto, a função de tokenização pode preservar as palavras interdependentes como, por exemplo, "cidade maravilhosa". O *token* "cidade", se isolado pode representar a semântica de um núcleo urbanístico qualquer, enquanto o *token* "maravilhosa" pode representar algo de conotação positiva. Porém, se juntas (*token* "cidade_maravilhosa") referem-se à cidade do Rio de Janeiro/Brasil.

2.5.2. Remoção de *stopwords*

Stopwords são palavras que aparecem com muita frequência no texto, como por exemplo, preposições, artigos e pronomes. Cada idioma deve possuir sua lista de *stopwords*. Estas palavras, em geral, devem ser desconsideradas por não agregar valor semântico ao texto e, portanto, são irrelevantes para qualquer domínio. Entretanto, neste trabalho, as *stopwords* são consideradas para o cálculo dos pesos (seção 2.3.2) nos casos em que haja casamento entre os trigramas (extraídos das discussões) em dados abertos. Por exemplo, o trigrama "rio de janeiro", deve manter a stopword "de" para preservar a semântica do termo composto (*token* "rio_de_janeiro"), já que esta encontra-se disponível em dados abertos.

2.5.3. Radicalização

A radicalização (ou *stemming*) é o processo de remoção dos sufixos das palavras, a fim de reduzi-las à sua forma raiz. Esse processo, também chamado de stemização (neologismo), é importante, pois a significação da palavra está relacionada ao seu

radical e deve ser considerado um mesmo termo para a contagem dos pesos (seção 2.3.2). Por exemplo, os *tokens* "computadores" e "computar", após aplicação da radicalização, possuem o mesmo radical "comput" [34].

2.5.4. Etiquetagem

A etiquetagem de classe gramatical (*Part of Speech Tagging – POST*) é o processo que associa o rótulo da função (ou classe gramatical) de cada palavra encontrada no texto (JURAFSKY e MARTIN, 2000). Além dos rótulos morfológicos (como substantivo, adjetivo, preposição etc.), também poderão ser etiquetadas as funções sintáticas das palavras (sujeito, predicado, objeto direto etc.) (BICK, 1998) ou funções semânticas das palavras, indicando seus papéis e informações sobre significado (VIEIRA, 2000). Normalmente, o processo é baseado em modelos dependentes de idioma, que são treinados a partir de *corpus* textuais anotados (MANNING e SCHÜTZE, 1999).

2.5.5. Desambiguação de Significado das Palavras

O objetivo da Desambiguação de Significado das Palavras (*Word sense disambiguation*) é determinar o significado das palavras no seu contexto de uso. O significado atribuído a uma palavra pode ser determinado, com certo grau de incerteza, pelo uso de técnicas computacionais que podem utilizar recursos estruturados (thesauros, dicionários, ontologia) ou desestruturados (*corpus* anotado) (NAVIGLI, 2009).

2.5.6. Seleção de Bigramas e Trigramas

Um n-grama é uma unidade linguística representada por uma palavra ou uma sequência de *n* palavras adjacentes. Nomeia-se "unigrama" palavras unitárias (*n*=1),

[34] Aplicação do algoritmo de *stemming* ORENGO

“bigrama” as sequências de duas palavras (*n*=2) e “trigrama” as sequências de três palavras (*n*=3). Por exemplo, a expressão “*software* colaborativo” é um bigrama e a expressão “caso de uso” é um trigrama.

A identificação dos bigramas e trigramas, adotados neste trabalho, pode ser realizada por intermédio de análise estatística (coocorrência), análise linguística, análise terminológica com auxílio de vocabulário controlado ou combinação dos métodos (CONRADO *et* al., 2009). Para todos os tipos de análise, as s*topwords* devem ser mantidas, pois cada combinação de palavras tem uma semântica específica associada.

Bigramas e trigramas são menos propensos a ambiguidade do que os unigramas e sua frequência no texto pode apontar possíveis conceitos de domínio (BEKKERMAN e ALLAN, 2004). No entanto, essa afirmação não garante que estas palavras compostas representem um conceito no domínio estudado (MANNING e SCHÜTZE, 1999). Usualmente, os métodos de extração automática de informação em base textual utilizam bigramas e trigramas, pela facilidade de extração e, principalmente, por serem menos propensos a ambiguidades que os unigramas (ESTOPÀ BAGOT, 1999) (VENTURA, 2008) (MOURA *et* al., 2008).

2.6. Agentes

Adotou-se a tecnologia de agentes para as tarefas de extração de informação, processamento linguístico e geração do contexto enriquecido. Também se adotou o conceito de gerenciamento indireto sob o ponto de vista da interação humano computador, para monitorar os eventos da interface e auxiliar o acesso ao sistema de informação proposto.

A definição do termo agente é muito ampla. Existem muitas propostas de taxonomias para esta área de pesquisa, mas não há consenso. O dicionário Aurélio

define agente como aquele que opera, ou aquele que age. No mundo da computação existem muitas outras definições, que variam de acordo com o ponto de vista e escopo de utilização (FRANKLIN e GRAESSER 1996). Tanenbaum e Steen (2002) definem agentes de *software*, no contexto de sistemas distribuídos, como um processo autônomo capaz de reagir ou iniciar mudanças no ambiente, possivelmente em colaboração com os usuários e outros agentes. Para Russell e Norvig (1995) "um agente é qualquer coisa situada em um ambiente que pode perceber este ambiente por intermédio de sensores e agir nesse mesmo ambiente por intermédio de atuadores". Portanto, a característica que torna um agente mais do que apenas um processo é a sua capacidade de agir por conta própria e, em particular, tomar a iniciativa sempre que necessário. Breitman (2005) consolida as seguintes características essenciais para agentes, aceitas, tanto pela FIPA (*Foundation for Intelligent Physical Agents)* quanto pela OMG (*Object Management Group*)[35]: (i) autonomia – capacidade de agir por conta própria, independente de controles externos, graças ao seu estado, que é encapsulado, (ii) interatividade – capacidade de comunicação com outras entidades e (iii) adaptabilidade – capacidade de reagir a estímulos, usando-os como forma de aprendizado e evolução.

Dois ou mais agentes atuando no mesmo meio formam um sistema multiagente (SMA). Agentes de um sistema multiagente devem ser sociáveis e colaborar mesmo se dispostos em ambientes heterogêneos e plataformas distribuídas (LIZOTTE e MOULIN, 1990). Mecanismos eficientes para comunicação entre agentes, que operem sobre algum tipo de linguagem padronizada, são mandatórios em um sistema multiagente (HÜBNER e SICHMAN, 2003). Segundo a OMG, para que um sistema seja considerado multiagente (SMA), devem existir mecanismos de coordenação entre os agentes, além das características essenciais (WOOLDRIDGE e JENNINGS, 1995). Ainda segundo

[35] http://www.omg.org/

Berners-Lee (2001), a *Web* Semântica só poderá atingir sua plenitude com a utilização de agentes.

2.6.1. Agentes Assistentes

Diversos trabalhos sobre agentes assistentes (ou agentes de interação) têm sido propostos para ajudar usuários a lidar com tarefas cotidianas ou complexas (PREECE *et* al., 2005), fornecendo-lhes informações de maneira simplificada (NORMAN, 1997). Originalmente vislumbrado por Nicholas Negroponte (1970), estes trabalhos visam diminuir a sobrecarga de informação e facilitar a vida do usuário (MAES, 1994).

Segundo Maes (1995) os agentes assistentes podem auxiliar os usuários de diversas maneiras, tais como: (i) executar tarefas, (ii) ensinar algo, (iii) viabilizar a colaboração entre usuários e (iv) monitorar eventos e procedimentos. Esses agentes podem executar tarefas complexas em *background*, que demandem tempo até serem concluídas. Portanto, este agente deve continuar atento a novas solicitações de tarefas e respondendo àquelas solicitações tão logo quanto possível.

2.6.2. Comunicação entre agentes

Para que a comunicação entre agentes seja possível, faz-se necessário a padronização de uma linguagem de comunicação entre agentes que formalize o intercâmbio de mensagens, tais como KQML [36] (*Knowledge Query Manipulation Language*) e FIPA-ACL [37] (*Agent Communication Language*). KQML foi o primeiro modelo para comunicação entre agentes, proposto pelo Departamento de Defesa norte americano, e define um protocolo com mensagens bem formatadas cada qual com seu significado específico (FININ e LABROU, 1997). Já a linguagem ACL é desenvolvida

[36] http://www.csee.umbc.edu/csee/research/kqml/
[37] http://www.fipa.org/repository/aclspecs.html

e mantida pela FIPA, que é uma organização que visa promover a interoperabilidade entre agentes heterogêneos a partir da especificação e padronização de protocolos para troca de mensagens, baseados na teoria da comunicação por atos de fala (FIPA, 2002).

2.6.3. *Framework* para Construção de Agentes

Segundo Johnson e Foote (1988), um *framework* é "um conjunto de classes que incorporam um projeto abstrato de soluções para uma família de problemas relacionados e possibilita reutilizações em uma granularidade maior do que classes". Existem diversos *frameworks* que se propõem a fornecer recursos de programação úteis à construção de sistemas multiagente. Dentre os mais populares estão o *JACK*[38], *Jadex*[39], *Jason*[40] e *Jade*[41] (adotado neste trabalho).

Estes *frameworks* aliados à engenharia de *software* orientada a agentes apresentam vantagens como a modularização, reusabilidade e facilidade de evolução do sistema. Essas vantagens são muito importantes, visto que podem reduzir custos e tempo de desenvolvimento (BITTENCOURT, 2006).

O *Jade* (*Java Agent Development Framework*) é um *framework* que provê um ambiente oportuno para a construção de sistemas baseados em agentes, de acordo com as especificações da FIPA. Cada instância do *Jade* (denominada *Container*) responsabiliza-se por um conjunto de agentes em uma Plataforma (*Platform*). Por sua vez, cada plataforma possui exatamente um *Container* principal denominado "*Main Container*". Na inicialização do "*Main Container*", dois Agentes especiais denominados AMS e DF são criados automaticamente. Outros *Containers* podem associar-se a um *"Main container"*, inclusive se distribuídos em diferentes pontos na rede.

---

[38] http://www.agent-*software*.com.au/products/jack/
[39] http://jadex-agents.informatik.uni-hamburg.de/xwiki/bin/view/About/Overview
[40] http://jason.sourceforge.net/Jason/Jason.html
[41] http://jade.tilab.com/.

O DF (*Directory Facilitator*) é um agente especial da plataforma *Jade* que reside em um "*Main Container*" e que provê serviço de "páginas amarelas" de uma *Platform*, ou seja, provê apoio para as tarefas de registro e busca de serviços prestados por agentes em uma *Platform.*

O AMS (*Agent Management System*) é outro agente que reside no "*Main Container*", que conhece todos os agentes da plataforma, controla os nomes dos agentes (devem possuir um nome único na *Platform*), provê a criação (ou registro) e remoção de agentes da *Platform,* mesmo em contêineres remotos.

A Figura 7, extraída de (CAIRE, 2009) ilustra um exemplo para os conceitos *Container*, *Platform*, DF e AMS.

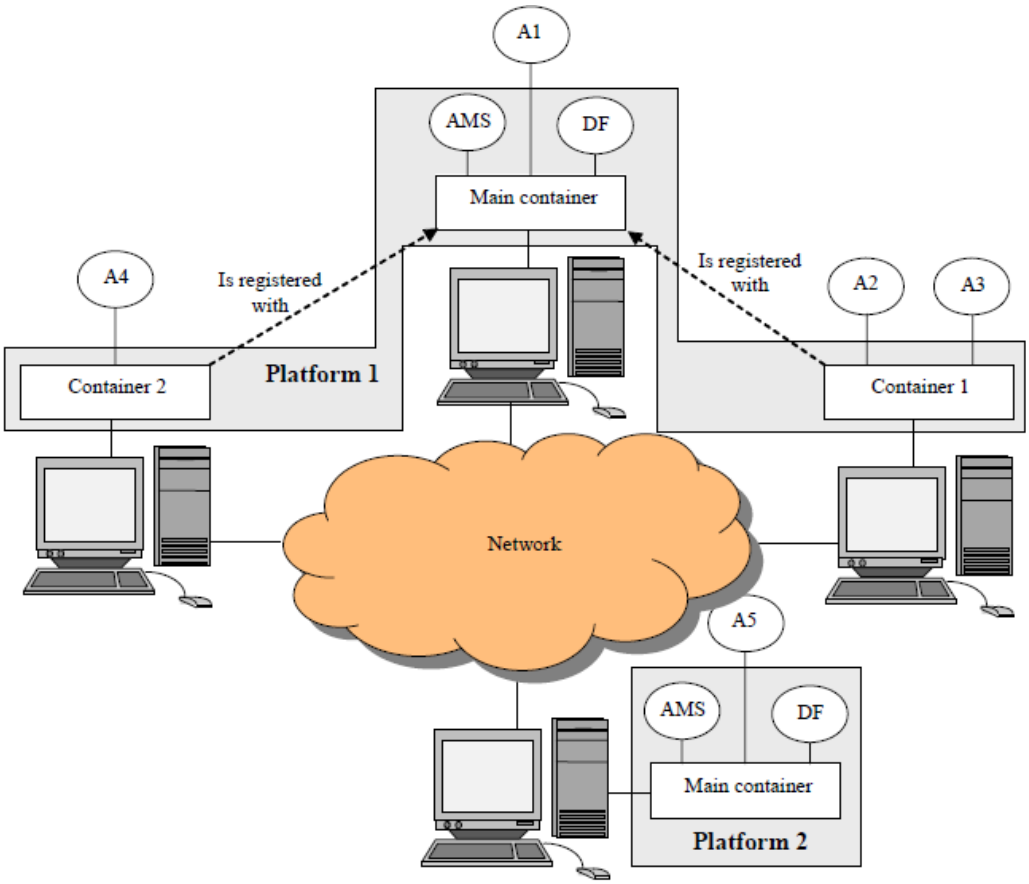


Figura 7 – *Containers* e *Platforms* do *Jade*

2.7. Trabalhos Relacionados

Esta dissertação relaciona-se com trabalhos que, independente de técnica, usam algum modelo de conhecimento para tornar a recuperação da informação sensível ao contexto. Informações de domínio do conhecimento, de processo de negócios, fornecidas pelo usuário (explicitamente) ou obtidas a partir de seus comportamentos (implicitamente) podem ser usadas para modelar o contexto. Esta modelagem pode exigir grande esforço humano, como a criação de ontologias por especialistas, preenchimento de preferências e marcação de documentos (modelagem manual) ou técnicas computacionais que supram (modelagem automática) ou minimizem (modelagem semiautomática) o esforço humano, como a análise de cliques, *corpus* textual, dados históricos, sensores ou outras maneiras ubíquas (BHOGAL *et* al., 2007) (CHANANA *et* al., 2004). Serão descritos alguns desses trabalhos, de modo especial, aqueles que empregam o *corpus* textual para a geração do modelo de contexto e, a partir deste modelo, propõe a recuperação da informação mais adequada às necessidades do usuário.

Prates e Siqueira (2011a, 2011b) propõem um método automático para apoiar a contextualização das atividades de busca na *Web*. Sua abordagem realiza a extração de conhecimento dependente de *corpus* (análise e processamento textual) para então utilizar técnica de expansão das consultas e as executa em motores de busca *Web*. A abordagem para a expansão de consulta assume que os termos mais frequentes em documentos que são representativos de um domínio têm maior probabilidade de ocorrer em sítios e documentos disponíveis na internet e são relevantes e relacionados a este domínio. Além disso, considera que um documento pode tratar de diversos assuntos e com isso, a criação da lista de termos, usada na expansão das consultas, faz uso de técnicas de segmentação de tópicos por assuntos antes da etapa de agrupamento

(*clustering*), aplicadas a um conjunto de arquivos de contexto representativos do domínio. Cada consulta expandida é executada no navegador *Web* e os resultados são apresentados ao usuário em cinco (5) abas distintas. Além da consulta original (aba 1), usada de *baseline* para a avaliação, a expansão pode considerar tanto os termos gerais do contexto (aba 2), quantas expansões com termos específicos de cada *cluster* de assunto (abas 3, 4 e 5). Assim o usuário pode visualizar e usar para a expansão os termos de todos os *clusters*, não só os termos do *cluster* que mais se relaciona aos termos da consulta. Kang *et* al. (2010) realizaram o agrupamento dos *x* primeiros documentos retornados pela consulta original e extraíram os termos mais relevantes de cada agrupamento. O usuário deve selecionar um dos *x* agrupamentos sugeridos, de acordo com o que julgar relevante. Os termos do agrupamento selecionado são somados aos termos da consulta original e o resultado da consulta expandida é apresentado ao usuário. Johnson *et* al. (2006) propõem uma aplicação para apoiar o processo de busca à *Web*. É realizado o *download* do conteúdo dos *x* primeiros documentos (*x=100*) retornados pela consulta original. Então, realiza-se a análise estatística para estes documentos à procura de bigramas e trigramas, que são exibidos ao usuário como sugestões para a expansão de consultas. Ambrósio *et* al. (2009) utilizaram técnicas de mineração de textos em um conjunto de documentos (apresentações) para recomendar documentos armazenados em um repositório e também sugerir a expansão da consulta a buscas na *Web*. Técnica de Mineração de Texto Educacional (*Educational Data Mining* [42] - EDM) é uma disciplina científica relativamente nova, que visa o desenvolvimento de métodos para descoberta de conhecimento e padrões em grandes coleções de dados educacionais (ROMERO *et* al., 2010) (SCHEUER e MCLAREN, 2012). Alguns trabalhos sobre técnicas de EDM podem ser encontradas em Baker e

[42] http://www.educationaldatamining.org/

Yacef (2009) e Romero e Ventura (2011). No entanto, esta abordagem baseia-se em grandes coleções oriundas de bases de dados de ambientes educacionais, que não é o caso deste trabalho, que se baseia em poucas mensagens para a definição do contexto. Os trabalhos apresentados nesse parágrafo são propostas de abordagem automática, que fornece um processamento mais rápido se comparado à modelagem especificada por especialistas do domínio de conhecimento para se criar o modelo de contexto.

Rensing *et* al. (2010) discutem a diferença entre os métodos baseados em conteúdo e sistemas de filtragem colaborativa para a recomendação de recursos de aprendizagem, propondo uma abordagem baseada em *tagging* colaborativo. Embora seja uma abordagem interessante, exige esforço dos participantes, que além de discutir e compartilhar conteúdo devem marcar os conteúdos nas discussões. Então, o conteúdo marcado nas discussões é processado, a fim de adquirir o contexto de aprendizagem. Zhuhadar e Nasraoui (2008) capturaram o contexto com o uso de taxonomias de domínio e perfis de usuário, e reclassificaram os resultados da pesquisa de acordo com a similaridade entre os termos contidos nos resultados da pesquisa e taxonomias. Paula (2010) propõe o uso de uma ontologia de domínio (ciência da computação) para apoiar a recuperação de informação. Usou-se o conhecimento desta ontologia e a técnica de expansão de consultas para recuperar documentos em uma coleção experimental composta por 889 artigos relacionados a subáreas de ciência da computação. Para a avaliação, utilizaram-se as medidas precisão, *recall* e medida F, e, com base nestas medidas, observou-se a superioridade das buscas expandidas com termos da ontologia. Este parágrafo identificou os trabalhos que usaram um modelo de representação independente de *corpus*, como o *feedback (*análise de relevância) ou modelagem manual de um domínio de conhecimento.

Capítulo 3 – Captura de Contexto a Partir de Discussões

Neste capítulo é apresentada a arquitetura proposta para captura de contexto a partir de discussões em redes sociais. Tem-se como objetivo a melhoria dos resultados das pesquisas com o uso da técnica de expansão de consultas. A motivação para essa proposta é apresentar resultados mais úteis em grupos de discussão, tais como ambientes de aprendizagem colaborativa.

3.1. Considerações Gerais

Existem diversas abordagens para a expansão de consultas, parte delas usa ontologias para modelagem manual do contexto. Estas ontologias são usadas para descrever o conhecimento a ser explorado. Modelar ontologias e realizar anotação semântica nas fontes de informação, possivelmente trará melhora na relevância na recuperação de informações, como consequência natural das boas anotações e boa modelagem conceitual e terminológica das ontologias. Resultados semanticamente anotados trazem resultados inequívocos. Porém, se essa fonte de informação for a *Web*, deve-se considerar a incompletude das anotações, já que seria muito difícil estimar o tempo e esforço gasto na modelagem, mesmo se restrito a um contexto específico.

Este trabalho optou pela criação (semi-)automática do contexto, devido ao esforço gasto por especialistas na criação manual, mesmo considerando que a criação automática pode levar a resultados incorretos.

3.2. Arquitetura Conceitual

Os principais buscadores do mundo, Google e Bing (ALEXA, 2012), priorizam, por padrão, documentos que contenham todas as palavras informadas na expressão de busca (GOOGLE, 2012) (BING, 2012). Assim, acrescentar palavras significativas ao contexto pode implicar em uma priorização melhor dos resultados, ou seja, mais de acordo com o contexto. A abordagem para expansão de consultas proposta nesse trabalho supõe que os termos mais frequentes em documentos e discussões que são representativos do domínio possuem maior probabilidade de ocorrer em sítios e documentos disponíveis na *Web*, que sejam relevantes e relacionados a este domínio (PRATES e SIQUEIRA, 2011a) (PRATES e SIQUEIRA, 2011b). A arquitetura desenvolvida (Figura 8) representa uma extensão do trabalho de Prates (2011).

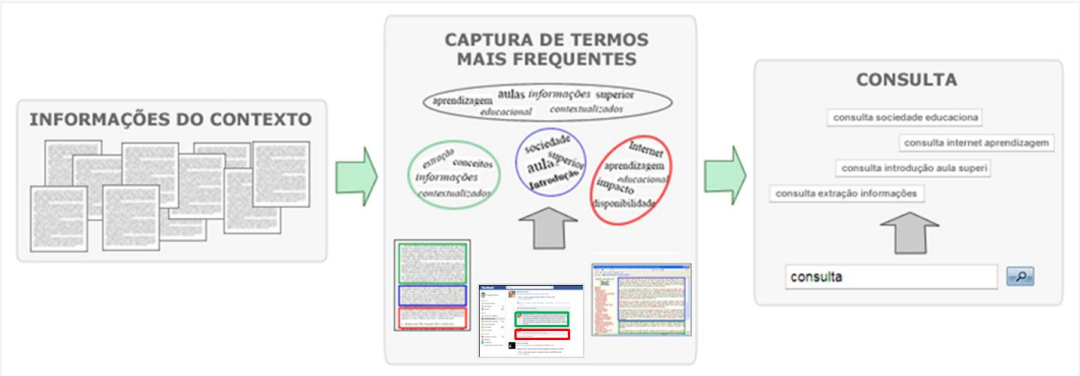


Figura 8 – Arquitetura Conceitual da Primeira Proposta

A manipulação de recursos textuais, como documentos (sítios, artigos, capítulos de livros, notas de aula, publicações em geral) foi estendida para considerar todas as mensagens trocadas (incluindo o conteúdo dos *links*) entre os participantes de um grupo em uma rede social (Figura 9) e, assim, obter o contexto do domínio e usá-lo para expandir as consultas dos usuários. Deste modo, seria possível obter recursos adicionais (de aprendizagem) a partir da *Web*, relacionados ao assunto do contexto (documentos ou discussões, no caso para apoiar o aprendizado). No entanto, esses documentos e

discussões podem tratar de assuntos diferentes e, nesse caso, a extração baseada na frequência dos termos a partir de todos os documentos pode misturar termos de diferentes assuntos à mesma consulta. Portanto, a tarefa de extração foi projetada para se obter os termos gerais do contexto e os termos específicos dos assuntos identificados.

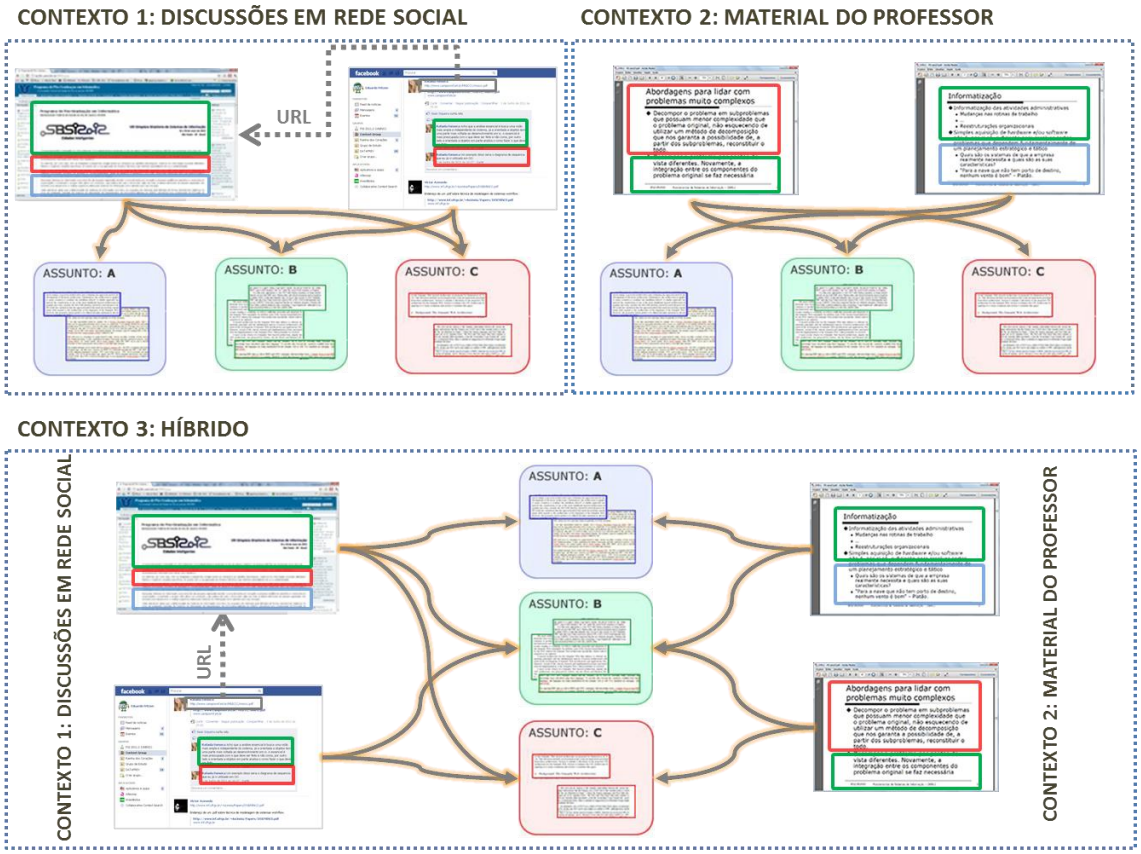


Figura 9 - Agrupamento de segmentos em assuntos

O conteúdo textual de todos os documentos e discussões do *corpus* são segmentados em tópicos, ou seja, trechos de documentos relacionados a um assunto específico são considerados documentos independentes. A segunda fase envolve o agrupamento desses segmentos em assuntos definidos de acordo com sua similaridade. A Figura 9 ilustra os segmentos (retângulos sobre o texto) e os três assuntos abordados (ASSUNTO A, B e C). Por fim, a extração de termos é realizada para cada assunto e dará origem a diferentes expansões de consulta (uma expansão de consulta diferente por assunto).

3.3. Arquitetura Lógica

A marcação em vermelho destaca as contribuições funcionais feitas por este trabalho, embora toda a arquitetura tenha sido remodelada, conforme (FRITZEN *et* al., 2011). A arquitetura está organizada em três módulos: Configuração da Base de Conhecimento, Extração de Informações e Busca. Será mostrada uma visão geral destes módulos, necessários ao entendimento deste trabalho.

Figura 10 – Arquitetura lógica

3.3.1. Módulo de Configuração da Base de Conhecimentos

Este módulo possui dois componentes: Seleção de Informações e Integração de Dados. Segundo Prates (2011), a camada Seleção de Informações pode contemplar "qualquer fonte de informação que contenha conteúdo textual e cuja informação represente assuntos abordados em um contexto de domínio". No âmbito deste trabalho, consideraram-

se anotações de aula fornecidas pelo professor e mensagens de discussão. Esses conteúdos devem ser acessados por intermédio da camada de integração de dados, que é capaz de ler e realizar a ligação de todo o tipo de informação disponível. A integração de dados permite combinar informações de documentos e *softwares* de comunicação.

3.3.2. Módulo de Extração de Informação

O objetivo deste módulo é identificar, para cada contexto, os principais termos de todas as informações obtidas a partir do módulo de Configuração da Base de Conhecimento e fornecer uma lista adicional de termos para o módulo de busca. Dois tipos de extração de termos são executados: os termos mais utilizados no contexto (extração de contexto) e termos específicos de cada assunto identificado no contexto (extração de assunto).

A extração dos termos específicos de cada assunto usa uma rotina de segmentação de texto para dividir o conteúdo da informação contextual em sentenças. A segmentação de texto é uma atividade de processamento de linguagem natural que tem como objetivo identificar subtemas dentro de um documento, definindo os seus limites. Cada segmento pode ser visto como um documento menor e independente. Em seguida, o agrupamento é usado para unir todos os segmentos semelhantes em documentos definidos por assunto. Por fim, os pesos de todos os termos são calculados para cada assunto e os *x* termos com maior peso são extraídos, onde *x* é o número máximo de termos que podem ser utilizados para a expansão de consulta. Em ambas as situações faz-se necessário aplicar algumas atividades de preparação de texto antes da extração de termos: *tokenization*, limpeza das *stopwords* e *stemming* (MANNING *et* al., 2008).

3.3.3. Módulo de Busca

O módulo de busca recebe: um contexto de indicação de domínio em que a informação precisa ser aplicada e as palavras-chave para realizar a pesquisa na *Web*. A consulta resultante é expandida usando os termos extraídos no módulo de Extração de Informação e executada em um buscador *Web*.

3.4. Arquitetura Física

A Figura 11 mostra a arquitetura física de componentes e demonstra cada tecnologia empregada no sistema. No Módulo de Configuração da Base de Conhecimento, duas fontes de informação foram consideradas: Material de aula no formato de arquivos PDF fornecido pelo professor e os *links* e comentários do *Facebook*. A leitura de arquivos PDF foi implementada por intermédio da biblioteca de código aberto (*open source*) *Apache PDFBox*. Esta biblioteca permite a extração de conteúdo textual simples dos arquivos, ou seja, são excluídos elementos de formatação e imagens.

Esta proposta usa a plataforma de rede social *Facebook*, que foi escolhida pela sua facilidade de integração com as demais tecnologias adotadas (protocolo *Open Graph*[43]) e, principalmente, pela sua popularidade no país e no mundo (COMSCORE, 2011).

Para acessar os dados do *Facebook*, adotou-se a biblioteca *RestFB*, que implementa o padrão de acesso *Open Graph*. *RestFB* foi escolhida, pois é escrita na linguagem de programação *Java*, a mesma usada no desenvolvimento de toda a arquitetura e, embora não seja oficialmente suportada, é *open source* e bastante utilizada. Toda a comunicação entre a aplicação desenvolvida e o *Facebook* foi através do uso de uma especificação de *JSON (JavaScript Object Notation).*

[43] https://developers.facebook.com/docs/opengraph/

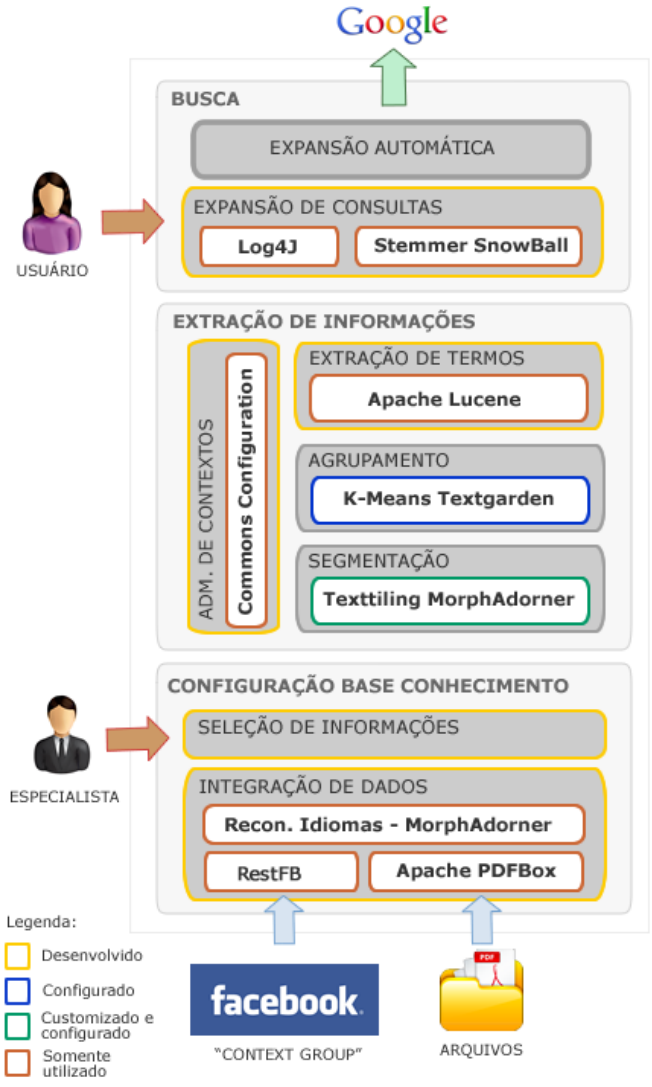


Figura 11 – Arquitetura física de componentes

MorphAdorner foi a biblioteca usada para a segmentação do texto e implementa o método *linear TextTiling* (HEARST, 1997), que consiste em gerar múltiplos segmentos de um texto. Cada segmento é um recorte do texto principal e origina um arquivo independente. *MorphAdorner* também é empregado para evitar a mistura de línguas diferentes no mesmo contexto, visto que isto poderia prejudicar a expansão das consultas.

Para o agrupamento dos segmentos, foi utilizada uma implementação do algoritmo de clusterização *k-Means* (MANNING *et* al., 2008) disponível na suíte de

softwares de mineração de texto chamada *Textgarden*[44]. A frequência dos termos é calculada para cada grupo de assuntos, de modo que os principais termos de cada grupo e de todo o contexto possam ser extraídos. A biblioteca *open source Apache Lucene*, uma *API* de indexação e pesquisa, foi usada para esse objetivo. O peso de cada termo em um documento (cada segmento foi considerado um documento) foi obtido com a aplicação do logaritmo na fórmula de frequência de termos, a fim de suavizar a influência desta frequência, tal como descrito em (MANNING *et* al., 2008). Finalmente, a biblioteca *Apache Log4j* foi utilizada para registrar as atividades de pesquisa e a avaliação feita na aplicação.

3.5. Abordagem Baseada em Agentes

Uma proposta de abordagem baseada em agentes para esta arquitetura é discutida em (FRITZEN *et* al., 2011). Essa proposta objetiva manter o contexto atualizado, por intermédio da distribuição de responsabilidades a agentes que verificam e tratam novas mensagens. Observou-se que, esta arquitetura pode ajudar a manter o contexto atualizado em ambientes dinâmicos, como grupos de discussão. Por outro lado, o tempo de processamento do contexto aferido foi expressivamente alto, fato atribuído à complexidade dos algoritmos de segmentação e agrupamento, o que poderia comprometer a utilidade das sugestões. Considerações de Desempenho desta arquitetura serão apresentadas na seção 4.5.1

[44] http://kt.ijs.si/dunja/textgarden/

3.6. Protótipo

A fim de testar a viabilidade das ideias preliminares, materializada na primeira arquitetura, um protótipo foi desenvolvido e integrado ao *Facebook Apps*. Cada consulta expandida foi executada em um motor de busca *Web* e os resultados foram apresentados ao usuário em abas diferentes (Figura 12). Uma expansão pode considerar o contexto geral (aba Contexto) e expansões com termos específicos de cada assunto (abas de Assuntos).

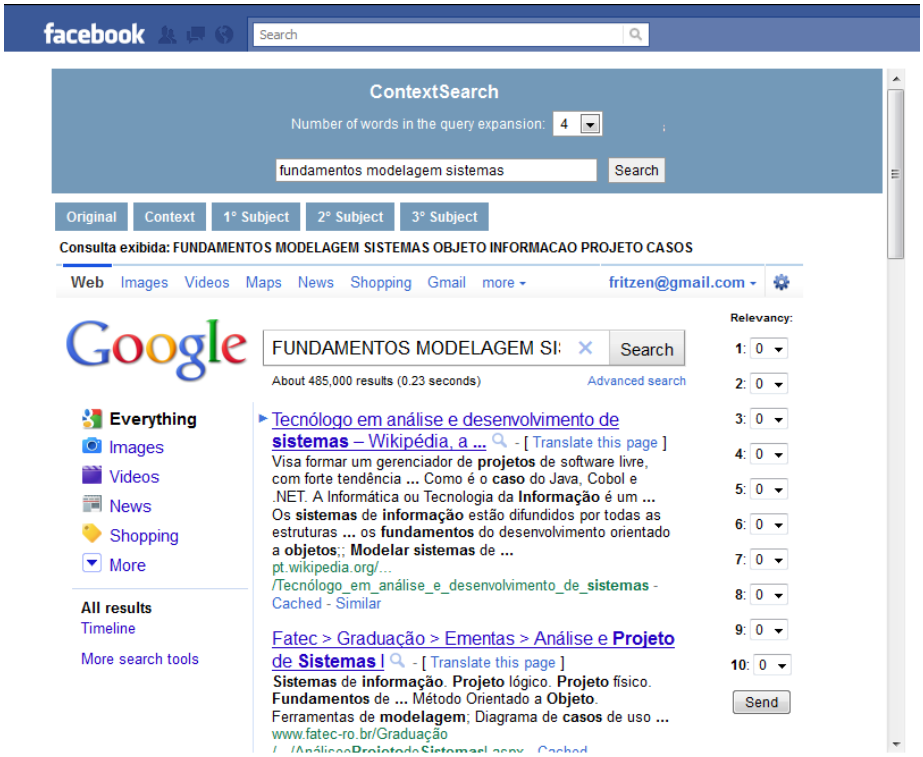


Figura 12 - Interface da expansão de busca considerando discussões

A modalidade adotada foi a expansão automática, que exige menos esforço do usuário para a execução das buscas. Além disso, observou-se que essa modalidade apresentou melhores resultados em (PRATES, 2011), possivelmente por, ao contrário da abordagem sugestão de termos, não permitir mistura entre termos de diferentes grupos de assuntos. Nessa modalidade, o usuário deve escolher em uma caixa de seleção o número de termos que devem ser incluídos à consulta original. Vale frisar que

esses termos são extraídos de maneiras distintas para cada aba de resultados, portanto são diferentes. Cada aba oferece como resultado um conjunto de dez documentos, que devem ser avaliados de acordo com sua relevância (0 = menos relevante, 4 = mais relevante). Existe uma caixa de seleção de relevância associada a cada um dos 10 documentos retornados em cada uma das cinco (5) abas (“consulta original”, “contexto”, “assunto 1”, “assunto 2” e “assunto 3”). Ao selecionar uma relevância para cada documento, deve-se clicar no botão de enviar para gravar a relevância dos documentos. Isso deve ser feito para cada aba e totaliza, portanto, 50 avaliações para cada consulta.

Capítulo 4 – Primeiro Estudo de Caso

Neste capítulo serão descritos o método da pesquisa, a dinâmica das atividades realizadas no primeiro estudo de caso, a caracterização dos participantes e os resultados de sua avaliação.

4.1. Metodologia

O método que mais se adéqua aos objetivos da pesquisa é o Estudo de Caso, pois, esta é uma pesquisa empírica, a qual busca encontrar relações entre causa e efeito que propiciem a compreensão de fenômenos. Yin (2005) afirma que "o estudo de caso é uma inquirição empírica que investiga um fenômeno contemporâneo dentro de um contexto da vida real, quando a fronteira entre o fenômeno e o contexto não é claramente evidente e onde múltiplas fontes de evidência são utilizadas". Este trabalho envolve a interação de alunos em um ambiente de aprendizagem colaborativa mediada por computador e posterior avaliação humana para os resultados da busca exibidos por um protótipo. A avaliação depende das características individuais de cada aluno, como seus de conhecimentos, capacidade de discernir, distinguir, colaborar etc.

Outro argumento em defesa do estudo de caso como método de pesquisa consiste no fato de que, embora a área de recuperação de informação (RI) forneça *software* para promover a avaliação automática, os textos submetidos a eles devem ser coleções

anotadas, o que exige tempo e esforço de pesquisa. Nesse âmbito, TRECs[45] (*Text REtrieval Conferences*) são conferências que visam avançar a pesquisa em RI. Para tanto disponibilizam grandes coleções de texto de referência anotadas, nas mais diversas áreas, para que os interessados possam usá-las em suas pesquisas e debater resultados homegenizados (mesma fonte de informação).

Contudo, recuperação de informação na *Web* é diferente de recuperação em textos anotados (dinamismo, ausência de um vocabulário controlado e a heterogeneidade dos tipos de documentos) (KOBAYASHI e TAKEDA, 2000) (KELLY, 2009). Como o objetivo da dissertação é encontrar documentos na *Web*, estes documentos, na sua grande maioria, apresentam-se em formato desestruturado e, portanto, não podem ser submetidos a tais *softwares* de análise. Além disso, a presente abordagem se baseia em buscas adaptativas, que utilizam informações do contexto para geração da recomendação de termos, com o agravante do dinamismo em que o contexto é gerado. O uso de métodos automatizados para a avaliação de buscadores adaptativos é um problema de pesquisa em aberto (VOORHEES, 2008).

O estudo de caso único foi conduzido em um ambiente acadêmico (UNIRIO) e teve a participação de alunos na utilização de um protótipo. O estudo de caso comparou a relevância da busca original (sem expansão, usada como controle) com a busca expandida.

4.2. Preparação do Ambiente

Esta dissertação adota plataformas de redes sociais para suportar tarefas de comunicação entre alunos e professores, conforme a literatura discutida na seção 1.4. Dentre os possíveis sítios de redes sociais, o *Facebook* foi escolhido por ser considerado

[45] http://trec.nist.gov/

um ambiente confortável aos participantes do estudo, visto que é utilizado pela maioria dos mesmos (COMSCORE, 2011).

Embora muitas linhas de pesquisa no campo da educação estão adotando plataformas de redes sociais para fins de aprendizagem em suas propostas, é válido observar que o uso do *Facebook* para fins educativos não o transforma em uma plataforma de aprendizagem. Faltam certos recursos ao *Facebook* e às redes sociais em geral, para torná-las plataformas de aprendizagem colaborativa (*Learning Management Systems*), como, por exemplo, o *Amadeus*[46], o *Moodle*[47], o *TelEduc*[48] ou o *BlackBoard*[49] (WANG *et* al., 2011).

Um roteiro foi elaborado para o estudo de caso e disponibilizado aos alunos. O roteiro é um documento que contém a "regra do jogo" para a participação no estudo de caso. Além do roteiro, o pesquisador aproveitou a reunião dos alunos na aula para explicar a dinâmica das atividades e o objetivo do estudo.

Os alunos e o professor foram instruídos a participar de uma aula baseada em discussão, realizada em um grupo da plataforma de rede social *Facebook*. O roteiro definiu os objetivos desta pesquisa, o tema da discussão e a maneira de utilização do *Facebook* no estudo. Os alunos foram instruídos a interagir livremente com o ambiente de tutoria colaborativa (participar da discussão no *Facebook*), a fim de adquirir e trocar conhecimentos acerca do tema proposto. Usou-se a mediação do professor para fomentar as discussões e acompanhar a sua aderência ao tema proposto.

[46] http://amadeus.cin.ufpe.br/index.html/
[47] http://www.moodle.org.br/
[48] http://www.teleduc.org.br/
[49] http://www.blackboard.com/

4.3. Dados Coletados

A avaliação e os resultados foram trabalhados a partir da coleta dos seguintes dados: histórico de mensagens no grupo, *log* do sistema, avaliação das buscas e questionários.

O histórico das mensagens serviu para traçar o perfil de uso da aplicação e a cronologia em que os fatos se sucederam durante a realização das atividades. O *log* é extraído do protótipo e armazena diversas variáveis de uso do sistema pelos usuários. A avaliação é o procedimento de ponderação feito para cada documento *Web* retornado (item do resultado de busca na *Web*) e é medida na escala de cinco (5) pontos (0=sem relevância a 4=relevante).

O questionário foi utilizado com a finalidade de ser mais um instrumento de coleta de dados e deveria ser preenchido ao final da dinâmica. As perguntas foram organizadas em duas partes, com objetivos distintos. A primeira parte objetivou traçar o perfil dos participantes (por exemplo, se preferem trabalhar sozinhos ou em grupo) e a segunda parte avaliar qualitativamente o protótipo (*Collaborative Context Search* - CCS), procurando entender, dentre outras coisas, qual contexto foi mais útil à busca.

Foram realizadas perguntas fechadas de múltipla escolha e perguntas abertas. Usou-se vocabulário simples, sem termos técnicos, com o intuito de evitar confusão. Foi dada preferência a perguntas no estilo de múltipla escolha pela facilidade de codificação, tabulação e comparação dos resultados, o que permitiu a comparação das opiniões de maneira objetiva.

Para a maior parte das perguntas de múltipla escolha, afirmações foram feitas acerca do protótipo e as alternativas seguiram a escala de LIKERT (1932), que consiste em cinco alternativas: “Discordo totalmente”, “Discordo parcialmente”, “Não concordo nem discordo”, “Concordo parcialmente” e “Concordo totalmente”.

As perguntas abertas visaram aprofundar o entendimento em relação à visão e opinião dos alunos, que puderam aproveitar a liberdade para expressar seus sentimentos, comentários e sugestões sobre o protótipo. Todos os participantes desta pesquisa tiveram suas identidades mantidas em sigilo.

4.4. Realização do Estudo de Caso

O primeiro estudo de caso foi realizado em um cenário de aprendizagem (PRATES *et* al., 2011) (PRATES *et* al., 2012), em um laboratório de informática da UNIRIO, com alunos do curso de Bacharelado em Sistemas de Informação, matriculados na disciplina "Fundamentos de Sistemas de Informação", explorando o tópico de "Modelagem de Sistemas".

Figura 13 – "*Context Group*"

O tema escolhido da aula para a realização do estudo de caso foi "Tecnologias usadas na Modelagem de Sistemas de Informação". O tema foi abrangente, com o objetivo de incentivar as buscas e as discussões entre os participantes. Os assuntos

envolvidos com este tema já haviam sido alvo de aulas presenciais. A dinâmica foi dividida em três etapas sequenciais: (i) discussões na rede social; (ii) busca e avaliação dos resultados da busca no protótipo; (iii) preenchimento de questionário. Toda a atividade da aula foi realizada através da colaboração entre os alunos, em um grupo da plataforma de rede social *Facebook*, chamado *Context Group* (Figura 13).

Neste grupo, os alunos debateram sobre os diversos assuntos envolvidos e, posteriormente, nessa mesma rede social, utilizaram o protótipo *Collaborative Context Search*, integrado ao *Facebook Apps*. Os alunos usaram o protótipo para pesquisar sobre uma necessidade de informação acerca dos temas discutidos durante a aula. As fontes de informação utilizadas para capturar o contexto foram os *slides* de aula fornecidos pelo professor e a extração de todo o conteúdo textual das discussões, incluindo *links* para sítios externos presentes nas postagens.

Para este estudo de caso, foram realizados testes em três modalidades de contexto distintas: (i) Contexto 1, gerado a partir do conteúdo das discussões, (ii) Contexto 2, material de aula (*slides*) fornecidos pelo professor e (iii) Contexto 3, híbrido, ou seja, abrange os conteúdos do Contexto 1 e do Contexto 2. Cada um dos três contextos foi subdividido em quatro partes (abas/painéis de resultado): uma para os termos gerais do contexto (aba Contexto) e as demais apresentando termos específicos de cada assunto (abas de Assuntos), conforme detalhado na seção 3.6. Para cada aba do conjunto de resultados retornados (10 resultados por aba), o aluno precisou avaliá-lo e gravar seu grau de satisfação, que variou numa escala de cinco pontos. Cada aluno decidiu por ele mesmo os critérios de diferencial semântico entre os bipolares completamente satisfeito e não satisfeito (OSGOOD *et* al., 1957). Para não afetar a avaliação, os alunos foram instruídos a serem justos e imparciais, ou seja, deveriam considerar somente o conteúdo

dos sítios, relevando outros critérios na avaliação, como *layout*, usabilidade, tempo de acesso, experiências passadas etc.

Os alunos contaram com um tempo de aproximadamente uma hora para a realização dos debates no grupo e uma hora para a busca e avaliação no protótipo. Para a avaliação, foram instruídos a formular uma expressão de busca que atendesse a sua necessidade de informação, referente ao assunto de pesquisa discutido no grupo *"Context Group"*, e selecionar o número de termos que o protótipo iria incluir na consulta original. Permitiu-se que os alunos reexecutassem a consulta quantas vezes fosse necessário, testando diferentes quantidades de termos, até que desejassem iniciar a avaliação de relevância dos resultados (refinamento da pesquisa).

4.4.1. Perfil dos Participantes do Primeiro Estudo de Caso

O primeiro estudo de caso contou com a presença de 20 alunos do curso de graduação Bacharelado em Sistemas de Informação da UNIRIO, que debateram, cerca de uma hora, sobre assunto definido pelo professor em um grupo de rede social. Do total de participantes, 16 realizaram as avaliações propostas no protótipo e 15 responderam o questionário (4 do sexo feminino e 11 do sexo masculino). Os respondentes possuem idade entre 18 e 36 anos, média de 22,67 anos e mediana de 19 anos. Todos possuem computador, acesso a internet e perfil no *Facebook*. Perguntados sobre como preferem estudar, 53% responderam "Sozinho" e 47% "Em grupo". O motivo para estudar sozinho foi "Concentro-me e raciocino melhor sozinho(a)" enquanto "Em grupo" foi "Permite a partilha de conhecimentos e ideias com os colegas".

4.4.2. Métricas Utilizadas na Avaliação

Como as métricas clássicas precisão e cobertura não são adequadas para se avaliar as buscas na Web (GRIFFITHS E BROPHY, 2005), foram exploradas outras métricas de avaliação sensíveis à ponderação humana: (i) precisão total dos *x* primeiros resultados, (ii) comprimento da busca e (iii) correlação de *ranking* (CHIGNELL *et* al., 1999) (COOPER, 1968) (SU *et* al., 1998).

Tang e Sun (2003) aplicaram essas três medidas a quatro motores de busca (Google, AltaVista, Excitar e Metacrawler). O objetivo do estudo não foi determinar o melhor motor de busca, mas sim testar e investigar diversas medidas de desempenho propostas por diferentes pesquisadores aplicáveis a estes motores de busca *Web*. A investigação concluiu que as três métricas propostas (*first 20 full precision, search length e rank correlation)* são melhores que alternativas tradicionais, como precisão e *recall*, pela ênfase à qualidade do *ranking*. De fato, o julgamento binário de relevância tomado nas avaliações tradicionais não leva em consideração a possibilidade de ocorrência de níveis de relevância nos documentos.

A relevância deve ser analisada a partir das perspectivas do usuário e do sistema. Já as métricas para mensuração de desempenho, são categorizadas como objetivas ou subjetivas (KOWALSKI, 1997). A perspectiva do usuário produz uma medida subjetiva, baseada no julgamento cognitivo de um usuário individual. Somente os usuários podem julgar, de acordo com critérios pessoais, sobre a adequação e relevância de cada documento retornado em satisfazer sua necessidade de informação. O juízo de relevância é mensurável em um ponto no tempo, ou seja, os critérios de relevância podem mudar com base no tempo ou na situação (KOWALSKI, 1997).

Já a perspectiva do sistema produz uma medida objetiva, a qual se baseia nos valores derivados da operação do algoritmo que relaciona as palavras contidas na

expressão de busca do usuário a um conjunto de documentos (similaridade). Não há dependência do usuário para o cálculo de relevância. O valor de relevância calculado é utilizado para a priorização (ordenação) dos resultados que são entregues aos usuários (KOWALSKI, 1997).

Para cada métrica, os dados obtidos com a consulta original foram comparados aos obtidos nas consultas expandidas. Se o resultado da consulta original for melhor que o resultado da consulta expandida, então conta-se um ponto para a consulta original. De maneira análoga, se o resultado da consulta expandida for melhor que o resultado da consulta original, então conta-se um ponto para a consulta expandida. Os resultados, apresentados em percentuais, são resultantes do total de pontos obtidos na contagem para cada tipo de consulta dividido pelo número total de consultas. Como os empates não foram contados, os valores percentuais não totalizaram 100%.

Para o primeiro estudo de caso, que além da consulta original contou com outras quatro consultas expandidas, cada qual relacionada a um assunto diferente (seção 3.3), considerou-se superioridade da consulta original somente se esta apresentasse melhores resultados que todas as demais expansões.

4.4.2.1. Precisão Total dos *x* Primeiros Resultados (CHIGNELL *et* al., 1999)

De acordo com Chignell *et* al. (1999), a medida de precisão total dos *X* primeiros resultados, que visa calcular a quantidade de informações relevantes encontradas nos primeiros resultados, é realizada de acordo com a seguinte equação:

$$\text{Precisão total dos X primeiros resultados} = \frac{\sum_{i=1}^{x} relevancia_i}{Y}$$

Onde: relevância$_i$ corresponde à pontuação (medida na escala de cinco pontos) atribuída ao i-ésimo resultado pelos avaliadores, *Y* é o resultado da multiplicação entre o

número total de resultados considerados (*X*) e o valor máximo de relevância possível de ser atribuído para cada resultado (4, uma vez que os valores vão de 0 a 4). A Tabela 6 ilustra os valores praticados para o estudo de caso.

Tabela 1: Valores para o estudo de caso

Variável	Estudo de Caso 1
Relevância máxima	4
Valor de X	10
Valor de Y	40

Embora o valor original de X para esta métrica seja 20, houve necessidade de redução da quantidade para 10 resultados, em todas as guias de resultados, tal como adotado em (PRATES, 2011). A justificativa para a alteração do número de resultados é a observação do comportamento dos usuários ao realizar buscas na *Web*. Usuários acessam poucas páginas de resultado, frequentemente restringem-se aos cinco primeiros resultados (SPINK e JANSEN, 2004) e (JANSEN *et* al., 1998). Além disso, usaram-se os resultados da primeira página de resultados do Google, que exibe 10 resultados por padrão.

4.4.2.2. Comprimento da busca (COOPER, 1968)

A segunda métrica utilizada foi o comprimento da busca, que indica o esforço do usuário em examinar resultados de pouca relevância, até encontrar *x* documentos consecutivos e relevantes. Esta pesquisa adotou os parâmetros *x=2* e valor de relevância igual a três ou quatro (numa escala de zero a quatro), os mesmos adotados por Tang e Sun (2003). Apenas os 10 primeiros resultados (primeiro estudo de caso) serão considerados. Considerando que Tang e Sun (2003) usaram um conjunto composto por

20 resultados, essa redução torna a medida muito mais rigorosa e dá uma ênfase muito maior aos primeiros resultados.

4.4.2.3. Correlação de *ranking* (SU *et* al., 1998)

Esta última métrica busca estabelecer uma medida de avaliação que correlaciona os *ratings* de relevância (escala de 5 pontos) obtidos pela avaliação qualitativa de seres humanos e o *ranking* atribuído pelo motor de busca para a priorização dos *x* primeiros resultados. O objetivo desta métrica é avaliar quão perto a classificação do sistema de busca está da avaliação ideal feita pelo usuário. A eficiência do mecanismo de busca pode ser denotada pela correlação entre os *x* primeiros resultados da busca e a avaliação humana para cada um destes resultados. Quanto maior a correlação, melhor será a eficácia do sistema de busca.

Como não existe acesso às notas reais de classificação atribuídas pelo sistema (Google no caso deste trabalho), utilizou-se a posição do documento no conjunto-resposta para presumir sua pontuação. Quanto maior a proximidade do documento ao topo da lista, maior a sua pontuação. Empregou-se o coeficiente de correlação de Pearson (KENDALL e STUART, 1973) para o cálculo da correlação entre a matriz A, que representa as avaliações dos usuários e a matriz B, que representa a ponderação associada a sua posição no conjunto de resultados da busca. O resultado da correlação entre as variáveis é apresentado no intervalo [-1, +1], onde os resultados próximos de +1 representam correlação direta e -1 correlação inversa.

Para o primeiro estudo de caso, dez resultados foram considerados, para cada um dos cinco conjuntos de resultados. Conforme ilustra a Tabela 2, para os dois primeiros resultados de cada conjunto atribuiu-se a pontuação quatro (pontuação máxima atribuída pelo usuário a um documento), para o terceiro e quarto resultados pontuação três, para o quarto e quinto resultados pontuação dois, para o sétimo e oitavo resultados pontuação

um e para os dois últimos resultados pontuação zero (pontuação mínima atribuída pelo usuário a um documento).

Tabela 2: Matriz de Correlação Para o Estudo de Caso 1

Resultado	Valor correlação
1	4
2	4
3	3
4	3
5	2
6	2
7	1
8	1
9	0
10	0

4.5. Avaliação do Primeiro Estudo de Caso

No primeiro estudo de caso, a avaliação quantitativa consistiu em julgar a relevância dos 10 primeiros resultados de busca obtidos na consulta original (aba 1) e nas consultas expandidas (contexto geral (aba 2), assunto 1 (aba 3), assunto 2 (aba 4) e assunto 3 (aba 5)). Assim, realizaram-se 50 avaliações no total para cada necessidade de informação. A modalidade adotada foi a expansão automática dos *x* termos mais relevantes do contexto, onde *x >= 1 e x <=10.*

Para comparar as mensagens de discussão e materiais de aula, três cenários para a aquisição do contexto foram considerados: discussões no grupo, incluindo o conteúdo textual de *links* externos (Contexto 1), notas de aula, obtidas a partir de *slides* do professor (Contexto 2) e híbrido (Contexto 3). Cada aluno avaliou pelo menos um cenário. Para o Contexto 1 foram consideradas 8 avaliações totais (todas as 5 abas) e 3 parciais (pelo menos uma aba). Para o Contexto 2 foram consideradas 5 avaliações totais e duas parciais e para o Contexto 3 foram consideradas 5 avaliações totais e 5 parciais. Foram descartadas do estudo avaliações incompletas para um mesmo conjunto

(aba) de respostas. Uma avaliação completa foi descartada, pois a expressão de busca foi considerada irrelevante e fora do contexto (a expressão foi “THE MONDAY ADVENTURE PELUDO”).

4.5.1. Considerações de Desempenho

O conteúdo textual gerado, após a conclusão do estudo de caso, é composto por 53 postagens textuais, com 44 postagens de *links* externos (páginas da *Web* ou arquivos pdf), e 69 respostas, além de 231 *slides* divididos em sete (7) arquivos fornecidos pelo professor. O computador usado na avaliação possui processador Intel dual core 2GHZ, 128GB SSD HD, 2GB DDR2 RAM e 15MB de conexão com a internet. O tempo total para a construção dos três contextos foi de 32 minutos, e inclui o *download* e o tratamento de todos os arquivos textuais. Outro fator que influenciou diretamente o alto tempo para a geração do contexto é o fato de que tanto o algoritmo de segmentação, quanto o algoritmo de agrupamento possuem complexidade assintótica alta e que deve ser feita para todo o *corpus*, sempre que houver uma nova postagem no grupo de discussões. Nesse tempo estão inclusas também outras atividades de complexidade linear (ou menor), como a extração dos termos mais frequentes de cada agrupamento, reconhecimento de idiomas realizado para cada segmento etc. Esses resultados afastam os algoritmos empregados nesta arquitetura do requisito de construção do contexto em tempo real.

4.5.2. Análise Quantitativa

Os resultados (Tabela 3) são apresentados em percentagens e representam a quantidade de vezes que o resultado obtido em cada aba foi melhor que as outras abas. A expansão dos assuntos considera os resultados em que pelo menos uma das três abas de resultados da consulta expandida com termos dos assuntos, apresentou melhores

valores do que a consulta original na métrica em questão; Já "Qualquer expansão" permite comparar o resultado obtido com qualquer tipo de expansão (expansão do contexto geral ou de qualquer uma das três expansões de assunto) em relação à consulta original.

Tabela 3: Resultados das Três Métricas

Expansão / Métricas	Precisão total	Comprimento da Busca	Correlação de *Ranking*
Consulta original	27%	38%	38%
Expansão do contexto	42%	38%	19%
Expansão dos assuntos	38%	35%	46%
Qualquer expansão	73%	38%	62%

Houve uma melhoria de 73% na medida de precisão total quando considerada pelo menos uma das expansões, enquanto a consulta original apenas mostrou melhores resultados em 27% dos casos. A melhoria percentual foi observada na comparação dos resultados a partir da consulta original com os resultados da consulta expandida com termos gerais do contexto (expansão geral) (42%), e comparando os resultados a partir da consulta original com os resultados de consultas expandidas com termos específicos de indivíduos (expansão assunto) (38%).

Em relação à métrica comprimento de busca, os resultados foram muito parecidos para todas as consultas (38%, 38%, 35%, 38%).

Melhorias nos resultados também foram observadas na métrica de correlação de *ranking*. A consulta original apresentou melhores resultados em 38% dos casos. A expansão com os termos gerais (expansão de contexto) apresentou melhores resultados em apenas 19% dos casos. No entanto, a expansão com termos específicos dos assuntos apresentou resultados melhores em 46% dos casos. Considerando qualquer expansão, os resultados foram melhores em 62% dos casos.

Outros resultados interessantes foram obtidos a partir da métrica de precisão total, considerando os diferentes cenários (contexto gerados a partir de anotações de aula, discussões ou de ambos), conforme apresentado na Figura 14.

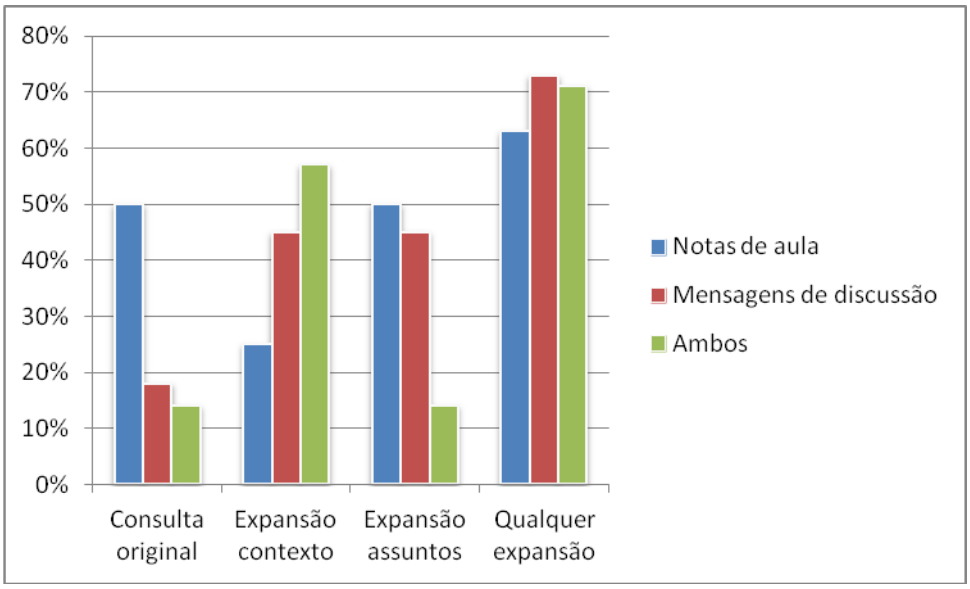


Figura 14 – Resultados da métrica de precisão total de acordo com o material usado na construção do contexto

As consultas originais apresentaram melhor desempenho com as notas de aula para a aquisição do contexto (50%) e com os termos específicos dos assuntos (50%), do que a expansão com os termos gerais do contexto (25%). Analisando os termos utilizados pelos alunos na expressão de busca, é possível entender a razão. As notas de aula versavam sobre conceitos gerais de Modelagem de Sistemas de Informação, mas as expressões de pesquisa observadas faziam referência a abordagens específicas de modelagem, que foram consideradas nas mensagens de discussão. Mesmo com esta diferença contextual, os resultados foram melhores em 63% dos casos quando se considera qualquer expansão da consulta.

Considerando-se as mensagens de discussão para as expansões de consulta, observa-se que as consultas originais proveram melhores resultados em 18% dos casos, enquanto as expansões (termos gerais e termos específicos dos assuntos) obtiveram 45%

das melhores avaliações. Considerando-se qualquer expansão, os resultados foram melhores em 73% dos casos.

Finalmente, a abordagem híbrida (considerando-se as anotações de aula e as mensagens de discussão do contexto) apresentou melhor resultado em 14% dos casos para a consulta original e expansão com termos específicos dos assuntos. A expansão com os termos gerais do contexto apresentou precisão total de 57%, e 71% ao considerar qualquer expansão. Em todas as métricas a expansão com termos gerais do contexto ou com termos dos assuntos apresentaram melhores resultados do que os obtidos com a consulta original.

4.5.3. Análise Qualitativa

O objetivo das questões apresentadas no questionário foi avaliar qualitativamente a relevância dos resultados sob a perspectiva dos alunos. Quinze avaliações foram coletadas a partir de questionários (alguns alunos preferiram não contribuir com os questionários).

Quando perguntados sobre a superioridade dos resultados de pesquisa obtidos com a expansão dos termos do contexto geral (segunda aba) em comparação aos resultados obtidos com a expansão dos termos extraídos dos assuntos (abas de assuntos), 58% dos alunos mantiveram-se neutros, mas 40% concordam total ou parcialmente, que os termos gerais do contexto apresentaram melhores resultados.

Eles também perceberam que as guias apresentaram diferentes visões sobre a mesma consulta (67% concordaram e 20% manteve-se neutra), mostrando a aplicabilidade da abordagem de segmentação e agrupamento proposta por Prates (2011).

De acordo com a Figura 15, ao serem questionados sobre qual contexto trouxe melhores resultados, 47% dos participantes apontaram o Contexto 01 (mensagens das

discussões), 33% escolheram o Contexto 02 (material de aula) e os 20% restantes responderam Contexto 03 (híbrido).

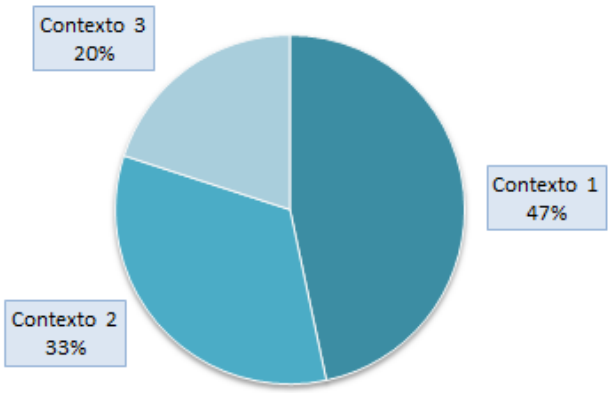


Figura 15 – Melhores resultados de acordo com questionário

Quando perguntados sobre o que eles acharam da experiência de aula colaborativa usando o *Facebook*, todos os alunos responderam que gostaram muito. Em geral, a impressão dos participantes em relação ao protótipo foi muito boa, porém relataram que a avaliação dos resultados foi uma tarefa exaustiva.

4.5.4. Considerações sobre a Avaliação

É importante considerar o comportamento dos usuários durante a realização do primeiro estudo de caso. Na dinâmica, os alunos deveriam interagir entre si em um grupo de rede social sobre o tema definido. As discussões eram livres e os alunos foram autorizados a realizar postagens textuais e *links*. Na prática, os alunos fizeram uso intensivo de motores de busca, com a intenção de encontrar *links* relacionados ao tema proposto para então compartilha-los no grupo. Poucas discussões (número de postagens textuais) foram observadas no início da dinâmica, até o professor intervir no grupo. Os melhores resultados de relevância encontrados no Contexto 1 (discussões no grupo, incluindo o conteúdo textual de *links* externos) estariam relacionados ao fato de que os *links* para documentos ou sítios que eles postaram no grupo foi relevante para eles.

Entretanto, os termos extraídos destes *links* levariam a documentos iguais ou muito semelhantes aos buscados previamente e esta poderia ser a origem dos resultados melhores.

4.6. Considerações Finais

Após a avaliação deste estudo de caso, evidenciaram-se requisitos que possivelmente melhorariam os resultados e ajudariam os alunos nas tarefas de busca a recursos na *Web*, visto que, os alunos aguardaram o final da etapa de discussão para depois usarem o protótipo de busca. Portanto, a primeira proposta não se adequa ao requisito de execução de buscas que possam sanar dúvidas durante a dinâmica e, assim, possa contribuir com as discussões. Com isso, os principais requisitos evidenciados foram: (i) permitir o uso do protótipo de busca durante o andamento das discussões no grupo e (ii) usar somente as mensagens para a modelagem do contexto (visto que esta apresentou bons resultados neste estudo de caso). Também se observou a necessidade de modificação da arquitetura e técnicas para adequar-se ao dinamismo da geração do contexto (o contexto cresce e molda-se de acordo com as discussões dos alunos).

Capítulo 5 – Enriquecimento de Termos das Discussões

Neste capítulo é apresentada a arquitetura proposta para enriquecimento do contexto a partir de dados abertos. O enriquecimento objetivou resolver o problema de pouca informação nas discussões, principalmente no início das mesmas (poucos termos). Tem-se como objetivo geral a melhoria dos resultados das pesquisas com o uso da técnica de expansão de consultas a partir de discussões, com termos cujo significado foi enriquecido com base no conteúdo de uma enciclopédia *Wiki*.

5.1. Justificativa

Esta nova proposta de arquitetura representa uma melhoria em relação à primeira proposta para permitir a execução de buscas durante as discussões. Esta melhoria relaciona-se diretamente ao tempo necessário para a execução do módulo de extração de informação de mensagens da primeira proposta de arquitetura. Na arquitetura anterior, novos documentos ou mensagens implicam em novas extrações de segmentos e agrupamentos, executados para todo o *corpus*. Isso demanda tempo e as consultas poderiam ser executadas em um contexto antigo. Isto pode ser um problema crítico para o caso de discussões, que resultam em um contexto mais dinâmico, como, por exemplo, permitir a execução de buscas durante dinâmicas de aprendizagem baseada em discussão.

Além do desempenho, outro fator preponderante à decisão de uma nova arquitetura foi a percepção da importância das discussões para a geração do contexto. A

avaliação da primeira arquitetura obteve bons resultados para o contexto gerado a partir de discussões, frente ao contexto gerado a partir de documentos (material de aula fornecido pelo professor). Entretanto, o comportamento dos alunos e a forma de extração do conteúdo das discussões devem ser considerados. Os alunos usaram motores de busca de maneira intensiva para pesquisar assuntos relacionados ao tema proposto, copiando e colando no grupo os *links* relevantes. Esse comportamento prolongou-se na maior parte da dinâmica e a discussão começou efetivamente somente após a interferência do professor. Como a extração considerou o conteúdo desses *links* para a construção do contexto, a quantidade e diversidade dos termos aferidos foram maiores do que os contidos no material fornecido pelo professor.

A partir do exposto até aqui, pensou-se em uma arquitetura para a geração do contexto somente a partir das discussões, no momento que elas ocorrem (em tempo real), para que a busca possa acontecer atualizada com as últimas postagens. Para a dinâmica, os alunos foram incitados ao debate. Embora não proibitivo, o envio de *links* foi desestimulado e desconsiderado para fins de extração de informação. Portanto, a proposta de arquitetura considerou: (i) o fato de não existir documentos (normalmente composto por muitas palavras), somente mensagens (normalmente composta por poucas palavras), (ii) não haver pré-processamento do contexto e (iii) o ambiente de discussão estar em constante modificação (maior dinamismo que a arquitetura anterior).

5.2. Arquitetura Conceitual

A arquitetura conceitual (Figura 16) deve atender à premissa de que, a partir de informações do contexto pobre (como as mensagens de discussão) e de uma base de conhecimento adicional, é possível obter o contexto de domínio enriquecido e, em seguida, usá-lo para expandir consultas *Web* do usuário a fim de obter recursos

relacionados ao assunto discutido. Muitas vezes encontrar os recursos desejados é uma tarefa difícil, porque, em geral, os usuários não expressam com precisão sua necessidade de informação e os motores de busca não se adaptam a diferentes interesses de usuários ou grupos de usuários. Para se evitar esse problema, o contexto deve evoluir de acordo com as novas necessidades de informação do usuário, sob a pena de perda de eficiência na busca e recuperação de informação. Para tanto, a atividade de enriquecimento do contexto deve ocorrer continuamente a partir da modificação nas informações do contexto e serve como insumo para as buscas dos usuários.

Figura 16 – Arquitetura Conceitual

5.3. Arquitetura Lógica

Para facilitar a explicação da arquitetura, dividiu-se a arquitetura lógica em dois subcomponentes: enriquecimento do contexto (seção 5.3.1) e processamento da consulta do usuário (seção 5.3.2).

5.3.1. Enriquecimento do Contexto

O papel do processamento do contexto é gerar o contexto de domínio enriquecido a partir da seleção de artigos de uma base de conhecimentos relacionados a mensagens de discussão. O contexto enriquecido é definido a partir da análise linguística e apoio semântico das discussões aliados a uma base conhecimento *Wiki*.

O contexto é gerado de acordo com as discussões, ou seja, somente mensagens são usadas para a construção do contexto. O conteúdo dos *links* publicados no grupo, abordado no primeiro estudo, não foi considerado neste. Para tanto, as seguintes etapas devem ser contempladas pela arquitetura (Figura 17).

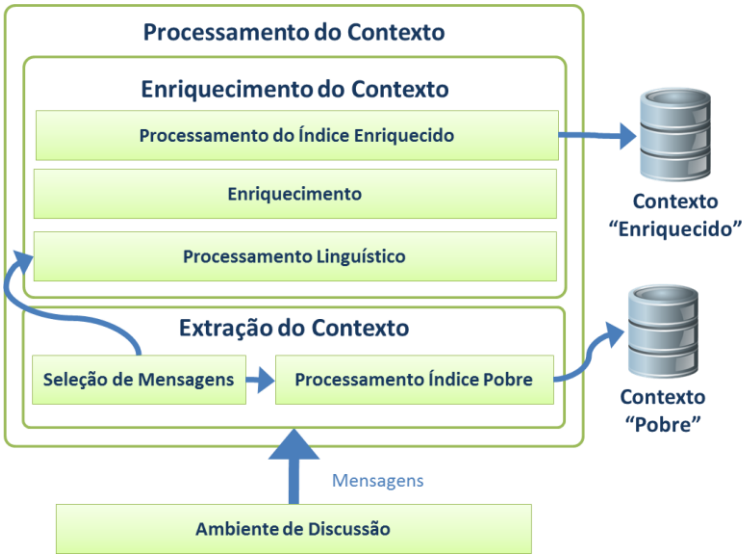


Figura 17 – Arquitetura Lógica para o Enriquecimento do Contexto

- Seleção de Mensagens – Responsável por obter as mensagens do grupo e replicá-las em uma base local denominada "Contexto Pobre" e também

repassá-las ao Processamento Linguístico para a geração do “Contexto Enriquecido”.

- Processamento Linguístico – Obter indícios nas discussões que se relacionem com artigos de uma enciclopédia *Wiki*.
- Enriquecimento – Com o apoio de uma ontologia que descreve a enciclopédia *Wiki*, obter o artigo da etapa anterior bem como artigos relacionados.
- Processamento do Índice – Organizar a informação obtida em um índice que garanta a rápida recuperação por semelhança entre o conteúdo do índice e uma expressão de busca.

5.3.1.1. Delimitação dos Dados Utilizados no Enriquecimento

A Figura 18 ilustra a abrangência dos artigos de uma enciclopédia colaborativa *online* considerada para a geração e uso do contexto enriquecido. Não houve necessidade de a arquitetura tratar o reconhecimento de idiomas, visto que, o idioma é um parâmetro para as discussões (o idioma do estudo de caso foi o português).

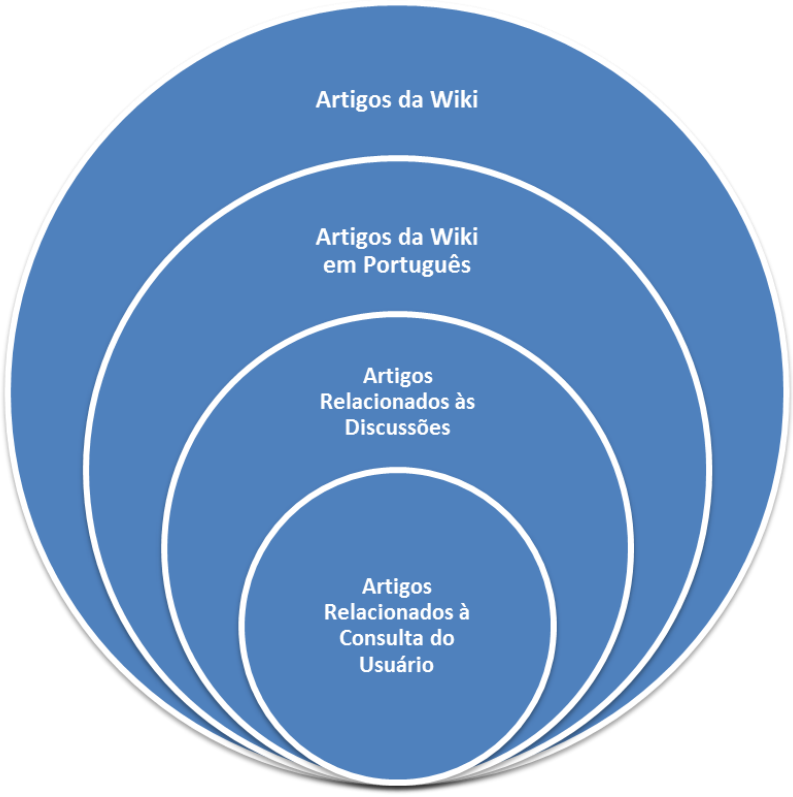


Figura 18 – Refinamento do Escopo

O subconjunto dos "artigos relacionados às discussões" é tratado pelo "processamento do contexto" e abrange os assuntos discutidos (interesse coletivo). Já o "processamento da consulta" aborda a porção dos "artigos relacionados à consulta do usuário" (expressão de busca), que representam o subconjunto dos "artigos relacionados às discussões" (interesse individual). Outra justificativa em se diminuir o escopo é que, segundo Steele (2001), motores de busca especializados (domínio específico) possuem índices menores e, portanto, mais gerenciáveis.

5.3.2. Processamento da Consulta

O processamento da consulta usa as mensagens e o contexto enriquecido (composto por artigos da *Wiki* obtidos no enriquecimento do contexto) para sugerir os termos que poderão ser usados pelo usuário para expandir a sua expressão de busca por documentos na *Web*. Os termos sugeridos, são exclusivamente aqueles que de alguma

maneira se relacionam à discussão. Portanto, a seleção dos termos para expansão é influenciada pela comunicação entre os usuários, ou seja, pelos eventos de colaboração realizados no ambiente de discussão, modificando-se com o tempo. Com isso, os termos mais relevantes no *tempo-1* talvez não sejam os mesmos no *tempo-2*, devido ao dinamismo do ambiente de discussões. A Figura 19 ilustra os componentes da arquitetura lógica para o processamento da consulta do usuário.

Primeiro, a expressão de busca é solicitada pelo usuário com o uso de sintaxe específica, conforme detalhado na seção 5.5 (protótipo). O componente de seleção de consultas deve identificar estas solicitações de busca, para então proceder à extração das sugestões dos termos candidatos à expansão da consulta. Três grupos de termos são extraídos de maneiras distintas. Os termos dos Grupos "A" e "B" são relacionados à expressão de busca, enquanto os termos do "Grupo C" independem da expressão de busca e representam os termos mais frequentes nas discussões no momento da busca. Ainda, os Grupos "A" e "B" são sugeridos a partir do "Contexto Enriquecido" e o "Grupo C" é sugerido a partir do "Contexto Pobre".

Figura 19 – Arquitetura Lógica para o Processamento da Consulta

O componente "Seleção Contexto Enriquecido", é responsável pela extração dos termos dos Grupos "A" e "B". A extração é feita a partir da análise da expressão de busca original e considera a similaridade semântica latente entre os termos da consulta e os documentos do contexto enriquecido. A Análise Semântica Latente é uma proposta para melhorar os problemas encontrados em outras abordagens de recuperação de informação baseada em palavras-chave, como o modelo de espaço vetorial (*VSM – Vector Space Model*). No modelo VSM, a matriz de documento possui, geralmente, dimensão elevada e esparsa, visto que cada palavra do *corpus* não ocorre em cada documento deste mesmo *corpus*. Ainda, segundo Rodrigues e Asnani (2010), a recuperação baseada em palavras-chave tem problemas em captar a estrutura semântica subjacente desejada e, com isso, torna-se vaga e suscetível a ruídos. Em outras palavras,

estas consultas podem retornar documentos irrelevantes ou excluir do conjunto de respostas documentos potencialmente relevantes penalizados por não conter pelo menos uma palavra-chave da consulta.

O "Grupo A" é composto por rótulos dos artigos de uma enciclopédia *Wiki* que apresentem em seu conteúdo maior grau de similaridade aos termos que compõem a expressão de busca. Consideraram-se os seis (6) rótulos de artigos mais relevantes. Poderiam ser compostos por mais de uma palavra, visto que os rótulos são obtidos a partir do título dos artigos da *Wiki*. O "Grupo B" é composto por termos com maior ocorrência, a partir dos conteúdos dos artigos retornados no "Grupo A". O objetivo deste grupo é possibilitar a sugestão de outros termos que sejam importantes, mas que não ocorram nos rótulos (títulos dos artigos). Em outras palavras, obter o conteúdo dentro dos documentos visa ampliar a completude terminológica, não se restringindo a apenas ao seu rótulo. Por exemplo, o artigo cujo rótulo é "Banco de Dados" [50] possui os termos {Atomicidade, Consistência, Isolamento e Durabilidade} em seu conteúdo. Para o "Grupo B", consideraram-se os seis (6) termos mais frequentes do conteúdo dos artigos do "Grupo A".

O componente "Consulta Contexto Pobre" é responsável pela extração dos termos do "Grupo C" e utiliza somente as mensagens de discussão para a extração dos termos e independe da consulta do usuário. Consideraram-se os seis (6) termos mais frequentes das discussões.

Já o componente de requisição é responsável por agrupar os termos extraídos e gerar o acesso à busca. Uma requisição é o resultado de uma solicitação de consulta que serve como linha de base para o protótipo de busca e avaliação dos resultados.

[50] http://pt.wikipedia.org/wiki/Banco_de_dados

Por fim, o componente de busca usa os dados consolidados na requisição para sugerir termos para a expansão de consultas. Adotou-se a modalidade de expansão de consulta interativa, visto que permite maior controle do usuário para a escolha dos termos que irão compor a expressão de busca (KANAAN *et* al., 2008). Segundo Carpineto e Romano (2012), a tarefa de compor a busca de maneira interativa por humanos pode evitar que os resultados da consulta se distanciem da intenção da busca (problema conhecido como *query drift*). Segundo Kelly *et* al. (2009), a sugestão de termos em uma consulta interativa pode ser especialmente importante nos casos em que os usuários efetuam busca em domínios sobre os quais eles têm pouco conhecimento ou familiaridade.

Além disso, ao contrário da segmentação e agrupamento relacionados à primeira proposta, todos os termos sugeridos são potencialmente relacionados a um mesmo assunto.

5.4. Arquitetura Física

A arquitetura física baseou-se nos ideais da arquitetura conceitual e lógica, e nas tecnologias da *Web* Semântica Social e de Agentes. Partiram-se das mensagens trocadas em uma rede social online e do conhecimento de uma enciclopédia *Wiki* (*Web* Social), considerando tecnologias e conceitos da *Web* Semântica, como o acesso à *DBpedia*, uma ontologia responsável por definir conceitos e relações entre os mesmos.

A enciclopédia colaborativa *Wiki* considerada neste trabalho foi a *Wikipédia*. O uso da *Wikipédia* é motivado pelo trabalho apresentado por Medelyan *et* al. (2009), o qual observa algumas pesquisas que usam as informações da *Wikipédia* para auxiliar diversas áreas como o processamento de linguagem natural, recuperação de informação, extração

de informação e construção de ontologias. Estes autores definem que a "*Wikipédia* é uma mina de ouro de informações".

Já a tecnologia de agentes foi adotada para fins de interoperabilidade e cooperação entre os módulos do sistema. Agentes na arquitetura proposta são pedaços de *software* que podem interagir com o ambiente (grupo de discussão) a fim de automatizar o processo de construção do contexto enriquecido e tratamento das solicitações de busca. Além disso, a distribuição de tarefas entre agentes e agentes em *hosts* é indispensável para a construção do contexto em tempo real.

A partir da arquitetura física, desenvolveu-se na linguagem de programação *Java* o protótipo de sistema de informação, baseado exclusivamente em componentes de uso gratuito e bibliotecas *open source*[51]. Foram pesquisadas as bibliotecas que melhor se enquadravam aos requisitos de desempenho relacionados ao tempo de resposta para as solicitações das consultas. O sistema desenvolvido deve ser eficiente (tempo), extensível (integração a outras redes sociais) e flexível (parametrizações).

Assim como na arquitetura lógica, optou-se por dividir a arquitetura física em duas partes: processamento do contexto (seção 5.4.1) e processamento das consultas (seção 5.4.2). Por fim, a seção 5.4.3 a apresenta uma proposta de solução multiagente para prover atualização contínua do contexto e atendimento às solicitações de busca para o ambiente de discussão.

5.4.1. Extração e Enriquecimento do Contexto

A proposta de arquitetura física para a geração do contexto a partir da extração de mensagens em grupos de discussão é apresentada na Figura 20. A partir das mensagens realizou-se o processamento de linguagem natural a fim de identificar as entidades

[51] http://opensource.org/

relacionadas e complementares ao contexto das discussões. Dois macrocomponentes podem ser identificados na proposta: “Extração do Contexto” e “Enriquecimento do Contexto”. A Figura 21 apresenta o diagrama de atividades para a arquitetura para o processamento das mensagens dos usuários.

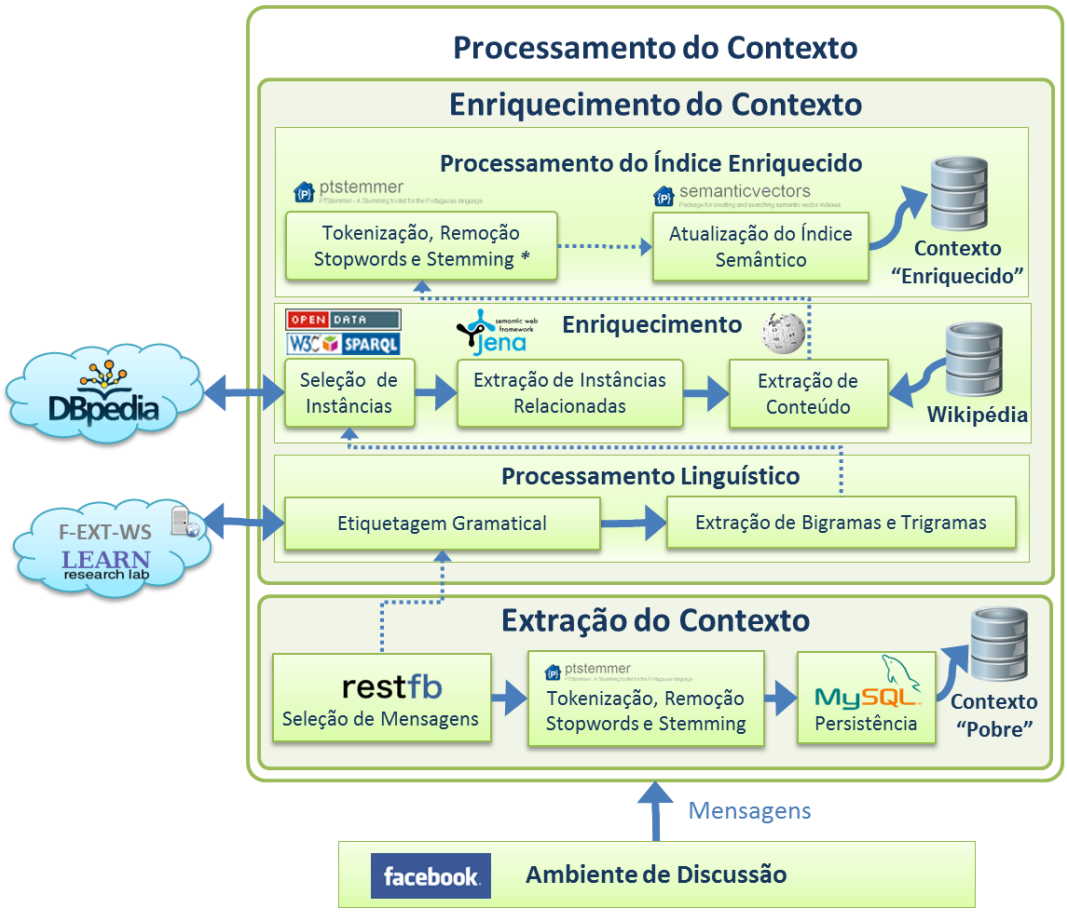


Figura 20 – Componentes da Arquitetura Física

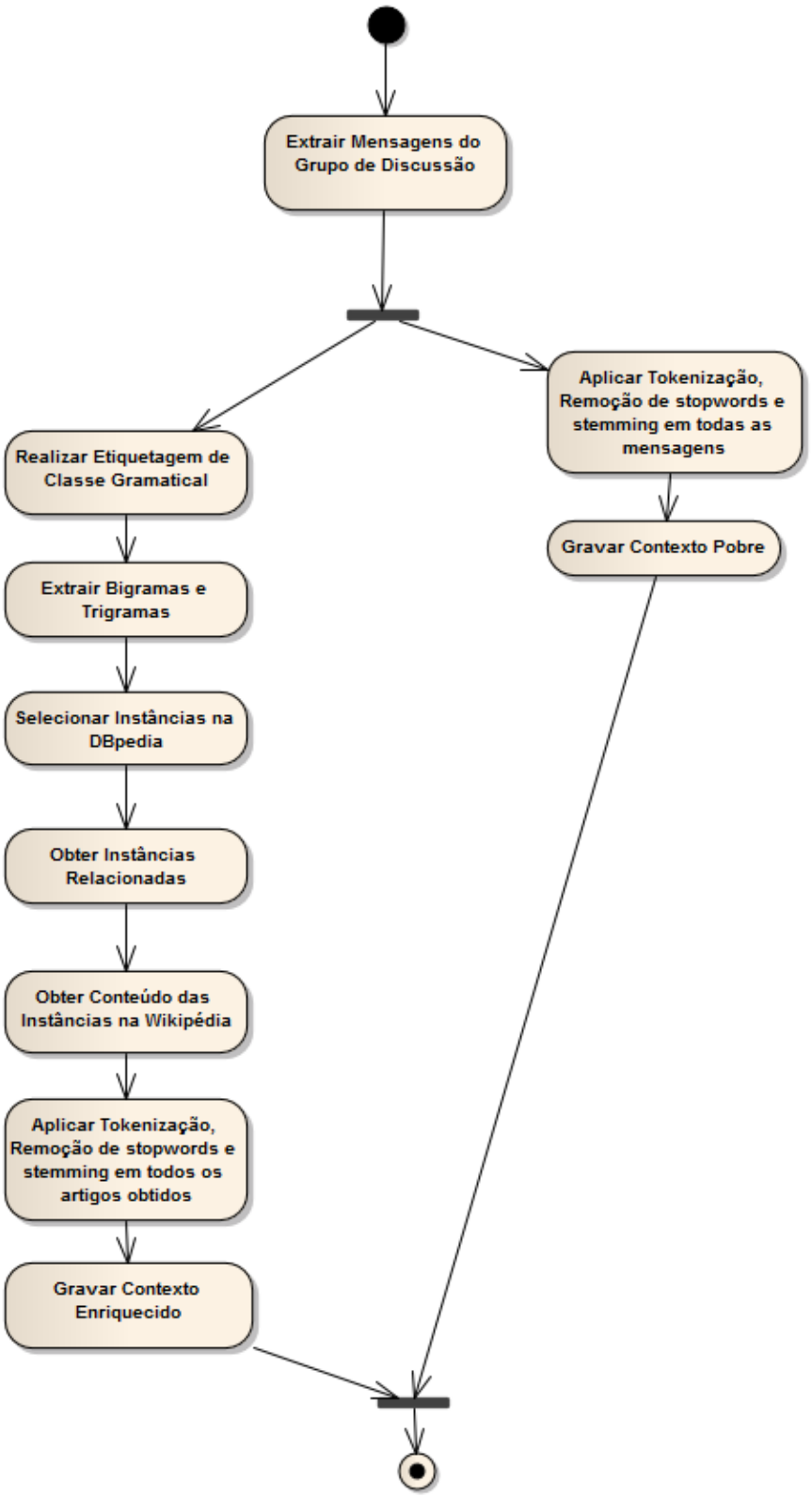


Figura 21 – Diagrama de Atividades para o Processamento das Mensagens

O objetivo da "Extração do Contexto" é obter o conteúdo textual das mensagens ou, especificamente, as postagens e os comentários de grupos do *Facebook*, incluindo

informações de autoria, data e hora. Para tanto, o componente "Seleção de Mensagens" adotou a biblioteca *RestFB*, que usa o *Open Graph* para a extração de informação desta rede social. Toda mensagem extraída é submetida à *tokenização*, remoção de *stopwords* e *stemming*, para então ser armazenada em um banco de dados relacional MySQL pelo componente de persistência, excluindo-se apenas as mensagens de consulta e oriundas do agente assistente. A "Tokenização, Remoção de *Stopwords* e *Stemming*" realiza as tarefas de, respectivamente, (i) quebra do conteúdo das mensagens em *tokens*, (ii) remoção de palavras com pouca importância e (iii) redução das palavras resultantes a sua forma raiz. O conteúdo destas mensagens origina o Contexto "Pobre", usado para que sejam extraídos termos do "Grupo C".

O componente "Seleção de Mensagens", além de fornecer conteúdo para a geração do Contexto "Pobre", também é responsável pela seleção de conteúdo textual para o enriquecimento do contexto, antes da etapa de *tokenização*, remoção de *stopwords* e *stemming* (ou seja, considerou-se a mensagem em seu estado natural). Para o enriquecimento do contexto, o componente de seleção de mensagens deve ainda selecionar apenas as mensagens que possuam duas ou mais palavras, desconsiderando-se os caracteres especiais, de pontuação e as *stopwords*, visto que, em etapas posteriores não se considerarão os unigramas, ou seja, não se desejou tratar as acepções das palavras para fins de desambiguação (*Word Sense Disambiguation*). Exemplos: ";-)", "<3", "da hora !", "legal !", "ok..." etc. O objetivo do "Enriquecimento do Contexto" é ampliar o contexto terminológico, enriquecendo-o a partir de dados abertos de uma enciclopédia colaborativa. Para tanto, três etapas devem ser cumpridas: processamento linguístico das mensagens, enriquecimento e processamento do índice semântico.

O "Processamento Linguístico" possui dois subcomponentes: "Etiquetagem Gramatical" e "Extração de Bigramas e Trigramas". A "Etiquetagem Gramatical" é

realizada para a identificação de bigramas e trigramas nas mensagens. Todas as mensagens são submetidas ao processamento de linguagem natural para definição da classe gramatical das palavras (*Part of speech / lexical class*). Para a geração da etiquetagem utilizou-se o serviço F-EXT-WS[52] treinado com o *corpus* em português Mac-Morpho (ALUISIO *et* al., 2003). Esse serviço é desenvolvido e mantido pelo LEARN[53] (*Algorithm Engineering and Machine Learning Laboratory*) da PUC-Rio e disponibilizado via *Web Service* (MOTTA *et* al., 2010). O F-EXT-WS foi escolhido pelos seguintes motivos: (i) Suporte ao idioma português; (ii) uso livre, mediante cadastro; (iii) acesso padronizado via protocolo WSDL, compatível com uma vasta gama de linguagens, dentre elas o *Java*; (iv) fácil acesso à documentação e a exemplos de uso; (v) suporte às funções de NLP *part-of-speech tagging*; (vi) uso corporativo e referenciado por uma série de artigos, teses e dissertações; (vii) boa acurácia; e (viii) boa performance (FERNANDES *et* al., 2009). Alternativas ao F-EXT-WS, tais como JtextPro[54], OpenNLP[55], LingPipe[56], LX-Suite[57] e NLTK[58] foram analisadas e rejeitadas por uma série de motivos, tais como não oferecer suporte ao idioma português, licenciamento, falta de documentação e exemplos de uso. A Figura 22 ilustra o resultado do processamento de etiquetagem gramatical, para a mensagem "Qual o momento certo para se fazer um plano de negócios?" submetida ao F-EXT-WS. Nesta figura é possível identificar o trigrama "plano de negócios", obtido pelo padrão {N, PREP, N}, ou seja, {SUBSTANTIVO, PREPOSIÇÃO, SUBSTANTIVO}. A lista completa de padrões gramaticais considerados é encontrada na seção 5.4.1.1.

[52] http://www.learn.inf.puc-rio.br/fextws/
[53] http://www.learn.inf.puc-rio.br/
[54] http://jtextpro.sourceforge.net/
[55] http://incubator.apache.org/opennlp/
[56] http://alias-i.com/lingpipe/
[57] http://lxcenter.di.fc.ul.pt/
[58] http://www.nltk.org/

Os bigramas e trigramas obtidos pelo componente "Extração de Bigramas e Trigramas" serviram para identificar instâncias de artigos na nuvem da *Linked Data*, em particular a *DBpedia*, usadas para a geração do contexto enriquecido. Adotou-se a abordagem linguística para a detecção dos bigramas e trigramas (análise individual, mensagem a mensagem, visto que as mensagens possuem poucas palavras). A abordagem estatística de coocorrência não foi considerada uma boa alternativa, pois, em geral exige um *corpus* maior (PEAT e WILLETT, 1991). Usou-se o conjunto dos bigramas e trigramas identificados como parâmetro para a formulação da expressão de busca por instâncias em dados abertos (extrair informações por significado, visto que bigramas e trigramas são menos propensos a ambiguidades).

```
1  [features = word, pos, np, ck, np, ne, ck, clause] [taggingTime=00:00:00]
2
3  Qual      PROSUB  0   B-NP 0   0   B-NP (S*
4  o         ART     I   B-NP I   0   B-NP *
5  momento   N       I   I-NP I   0   I-NP *
6  certo     ADJ     I   I-NP I   0   I-NP *
7  para      PREP    0   B-PP 0   0   B-PP *
8  se        PROPESS 0   B-NP 0   0   B-NP *
9  fazer     V       0   B-VP 0   0   B-VP *
10 um        ART     I   B-NP I   0   B-NP *
11 plano     N       I   I-NP I   0   I-NP *
12 de        PREP    I   B-PP I   0   B-PP *
13 negócios  N       I   B-NP I   0   B-NP *
14 ?         ?       0   0    0   0   0    *S)
```

Figura 22 – Resultado do Processamento de Etiquetagem Gramatical

O "Enriquecimento" possui três subcomponentes: "Seleção de Instâncias", "Extração de Instâncias Relacionadas" e "Extração de Conteúdo das Instâncias". A "Seleção de Instâncias" tem por objetivo buscar todos os indícios de bigramas e trigramas em uma ontologia de enciclopédia colaborativa – *DBpedia* – para obter o apoio semântico em um vocabulário controlado. A proposta de arquitetura defende o princípio de que bigramas e trigramas encontrados em mensagens de discussão são representativos do domínio discutido e, portanto, são candidatos em potencial a aparecer em rótulos de instâncias na *DBpedia*. Palavras simples foram desconsideradas da

solução na etapa de expansão de dados abertos, pois as relações entre as palavras simples possuem alto nível de ambiguidade como conotação (uso da palavra em sentido diferente do original), homônimos (significado diferente, mesma grafia) ou polissemia (muitos significados para a mesma palavra). A problemática que envolve o tratamento automático de ambiguidade (MCCARTHY et al. 2007) é um campo de estudo reconhecidamente complexo (SMEATON, 1997) (NAVIGLI, 2009). Além disso, observa-se que os bigramas e trigramas representam a maior parte dos artigos da *Wikipédia*. A Figura 23 ilustra a composição dos rótulos (títulos dos artigos) da língua portuguesa na *Wikipédia*: apenas 29% são unigramas, enquanto 71% são ($n > 1$)-gramas. Destes 71%, 68% correspondem a bigramas ou trigramas e 32% a ($n > 3$)-gramas.

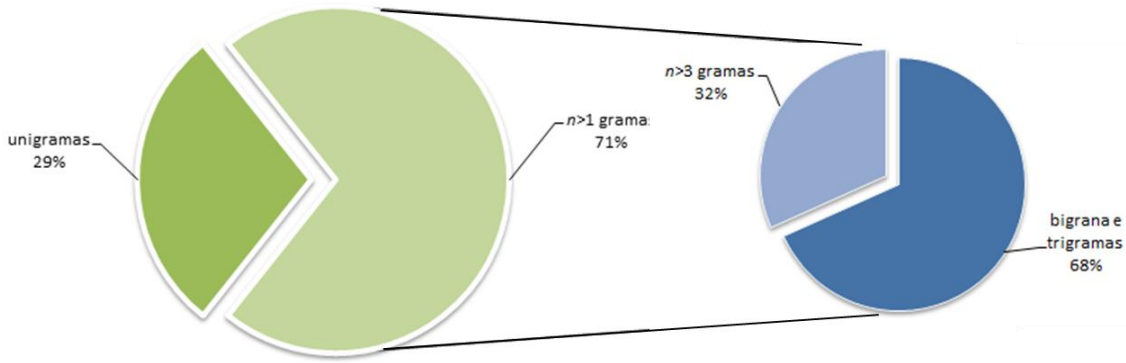


Figura 23 – Quantidade de N-gramas que compõem a *Wikipédia*

Para o casamento entre mensagens e instâncias da *DBpedia*, todos os substantivos e adjetivos dos bigramas e trigramas das mensagens foram submetidos ao processo de *stemming*, quando são adicionados caracteres coringa, substituindo os caracteres removidos pelo *stemming*. Por exemplo, o bigrama "Estruturas organizacionais" vira "estrut* organizac*", onde o coringa asterisco (*) pode corresponder a quaisquer sequências de caracteres. Nesse caso, os espaços e as *stopwords* são importantes e devem ocorrer na expressão de busca por instâncias. Isoladamente, "estrutura" e "organizacional" possuem um amplo leque semântico, dependente do contexto que foi

escrito, enquanto "estrutura organizacional", restringe radicalmente problemas de interpretação e ambiguidades. Com o apoio da biblioteca Jena [59], buscou-se o mapeamento entre bigramas e trigramas a instâncias da *DBpedia*. Para tanto, utilizou-se o rótulo *rdfs:label* e consultas na *DBpedia* a partir do SPARQL (WANG *et* al., 2007). Os rótulos *rdfs: label* dos recursos da *DBpedia* são criados a partir dos títulos das páginas da *Wikipédia*. As instâncias selecionadas são todas aquelas em que houver casamento (*matching*) entre os bigramas e trigramas encontrados no texto e os dados abertos, independente da linguagem, embora o componente de "Extração de Conteúdo das Instâncias" considere apenas aquelas relacionadas ao idioma informado como parâmetro para as discussões.

O componente de "Extração de Instâncias Relacionadas" deve agregar à base contextual enriquecida as instâncias relacionadas àquelas obtidas pelo componente "Seleção de Instâncias". Para tanto, utilizou-se a propriedade Dublin Core Subject (*dcterms:subject*) e o apoio da biblioteca Jena para o tratamento das triplas RDF. A Figura 26 ilustra este tratamento em um trecho de código RDF extraído da *DBpedia*. De acordo com a figura, o sujeito (*rdf:Description*) é "*rdf:about = Management_information_system*", o predicado (*dcterms:subject*) e o objeto "*rdf:resource=Category:Management_systems*" que é uma coleção de instâncias relacionadas ao sujeito.

[59] http://incubator.apache.org/jena/

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:dcterms="http://purl.org/dc/terms/">
  <rdf:Description rdf:about="http://dbpedia.org/resource/Management_information_system">
        <rdfs:comment xml:lang="pt">Sistema de informação de gestão ou sistema de ...</rdfs:comment>
        <rdfs:comment xml:lang="en">A management information system (MIS) is ...</rdfs:comment>
        <rdfs:comment xml:lang="es">Estos sistemas son el resultado de ...</rdfs:comment>
        <rdfs:comment xml:lang="de">Ein Management-Informationssystem ...</rdfs:comment>
        <rdfs:label xml:lang="pt">Sistema de informação de gestão</rdfs:label>
        <rdfs:label xml:lang="en">Management information system</rdfs:label>
        <rdfs:label xml:lang="de">Management-Informationssystem</rdfs:label>
        <rdfs:label xml:lang="es">Sistemas de información gerencial</rdfs:label>
        <dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:Management_systems" />
        <dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:Information_systems" />
        <dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:Information_technology_management" />
    </rdf:Description>
</rdf:RDF>
```

Figura 24 – Trecho Simplificado de Código RDF para a instância "Management_information_system.rdf"

Consideraram-se todas as instâncias obtidas pela relação de afinidade direta, ou seja, artigos de uma mesma categoria (*dcterms:subject*). Categoria[60] (por exemplo, *Category:Management_systems*) é um tipo especial de recurso da *DBpedia*, que é extraído a partir da classificação e agrupamento de artigos na *Wikipédia* que tratem sobre um mesmo assunto. Pode ser vistos como um conceito abstrato (*Concept*) que agrupa um conjunto de recursos (artigos) relacionados (MIRIZZI, 2010). Cada recurso do tipo *Category* na *DBpedia* associa-se a um *Concept* SKOS[61] (*Simple Knowledge Organization System*) por intermédio da propriedade *<rdf:type rdf:resource="skos:Concept">*, conforme exemplificado na Figura 25.

[60] http://en.wikipedia.org/wiki/Help:Category
[61] http://www.w3.org/2004/02/skos/

```xml
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:dcterms="http://purl.org/dc/terms/" >
  <rdf:Description rdf:about="http://dbpedia.org/resource/Workforce_modeling">
    <dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:Management_systems" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Management_information_system">
    <dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:Management_systems" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Executive_information_system">
    <dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:Management_systems" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Purchase_order_request">
    <dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:Management_systems" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Category:Management_systems">
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept" />
    <owl:sameAs rdf:resource="http://dbpedia.org/resource/Category:Management_systems" />
    <rdfs:label xml:lang="en">Management systems</rdfs:label>
    <skos:prefLabel xml:lang="en">Management systems</skos:prefLabel>
  </rdf:Description>
</rdf:RDF>
```

Legenda:

- Diferentes recursos (artigos) que compartilham uma mesma categoria (“Management systems”)
- Categoria “Management systems”
- Associação da categoria “Management systems a um conceito SKOS.

Figura 25 – Trecho de Código RDF para a instância “Category-Management_systems.rdf”

A Figura 26 apresenta a visualização na forma de um grafo dirigido de um trecho do código RDF apresentado na Figura 24. As arestas representam a ligação entre nós de recursos extraídos da *Wikipédia*. Estas ligações são identificadas por rótulos (predicados) que relacionam os nós sujeito e objeto. Esta representação gráfica é um modelo mais simples de ser visualizado por humanos e pode ser útil para a compreensão das relações.

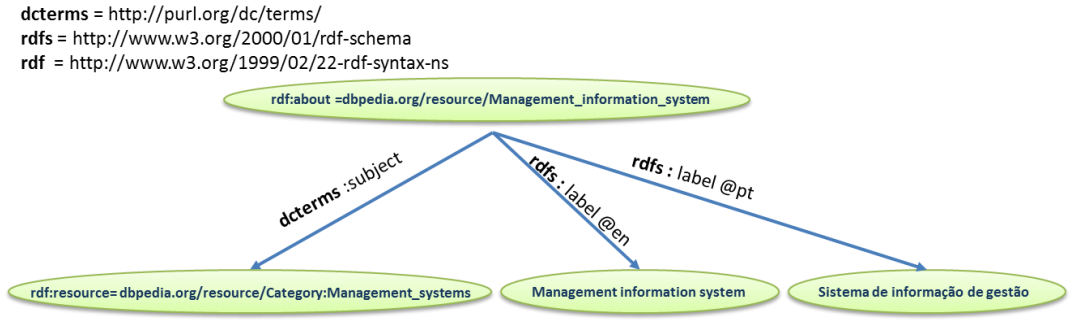


Figura 26 – Exemplo de Grafo RDF para a instância “Management_information_system”

O componente “Extração de Conteúdo das Instâncias” é responsável por obter o conteúdo na *Wikipédia* de todas as instâncias, tanto aquelas obtidas pela “Seleção de Instâncias”, quanto as obtidas pela “Extração de Instâncias Relacionadas”. Como a *DBpedia* não traz o conteúdo do artigo, a base da *Wikipédia* foi baixada (*dump*) e indexada em um repositório local com o apoio do Apache Lucene[62], uma biblioteca de indexação e busca textual escrita em *Java*. Os passos e a motivação para a indexação em um repositório local são apresentados na seção 5.4.1.2. Para todas as instâncias selecionadas, deve existir o recurso na língua definida para o contexto, ou seja, somente serão considerados para o contexto enriquecido os artigos relacionados que se enquadrem na linguagem da discussão.

Por fim, é realizado o “Processamento do Índice Semântico”, composto por dois subcomponentes, “Tokenização, Remoção *Stopwords* e *Stemming*” e “Atualização do Índice Semântico”. Nesta camada, a “Tokenização, Remoção de *Stopwords* e *Stemming*” modifica o conteúdo de instâncias dos artigos obtidos pela “Extração de Conteúdo” e realiza as tarefas de, respectivamente, (i) quebra do conteúdo dos artigos em *tokens*, (ii) remoção de palavras com pouca importância do texto e (iii) redução das palavras resultantes a sua forma raiz. No entanto, existe uma exceção à regra geral para os bigramas e trigramas. Estes serão unidos e tratados como um único *token*, no caso em que houver casamento entre estes termos compostos em dados abertos. Por exemplo, se o trigrama “rio de janeiro” ocorre em dados abertos, deve-se considerar o *token* “rio_de_janeiro” e, em seguida, ignorar a remoção da *stopword* (preposição “de”) e stemizar as demais palavras, resultando no *token* “rio_de_jan”. Já o tipo de stemização

[62] http://lucene.apache.org/java/docs/

adotado foi o ORENGO, realizado a partir do PTStemmer[63], uma biblioteca de stemização para o idioma português (ORENGO e HUYCK, 2001).

Por fim, a subtarefa de "Atualização do Índice Semântico" deve ocorrer para que o contexto enriquecido possa ser pesquisável por intermédio da técnica de análise semântica latente. Para tanto, usou-se a biblioteca *Semantic Vectors*[64] para explorar o espaço semântico textual e resolver problemas de relacionar a expressão de busca a um conjunto de documentos que compõem o contexto enriquecido. O *Semantic Vectors* usa os dados indexados no Lucene para a geração da matriz de termos e documentos.

5.4.1.1. Padrões Gramaticais Considerados

Este trabalho, assim como (ZAVAGLIA et al., 2007), utilizou a abordagem linguística e os padrões descritos na Tabela 4 para a identificação dos bigramas (por exemplo, padrão N + ADJ = {"livro didático", "sistema colaborativo", etc.}) e trigramas (por exemplo, padrão N + PREP + N = {"sistema de informação", "banco de dados", etc.}) nas mensagens. O significado para estes padrões são mostrados na Tabela 5.

Tabela 4: Padrões morfossintáticos utilizados

Bigramas
N + ADJ
ADJ + N
N + N
ADJ + ADJ

Trigramas
N + PREP + N
N + (PREP + ART) + N
N + PREP + ADJ
N + (PREP + ART) + ADJ
N + N + ADJ
N + ADJ + ADJ

Outras combinações, tais como PREP+ADV ("por meio", "por enquanto", "até aqui", "até hoje", etc.), PREP+ADJ ("de novo", "em comum", etc.) e PREP+N ("em relação", "por exemplo", "por causa" etc.) não foram utilizadas pois tem poucas

[63] http://code.google.com/p/ptstemmer/
[64] http://code.google.com/p/semanticvectors/

chances de representar um conceito e, portanto, não são bons candidatos a ocorrer em dados abertos.

Tabela 5: Abreviaturas das classes gramaticais utilizadas

Abreviatura	Significado
N	Substantivo Comum ou Próprio
ADJ	Adjetivo
PREP	Preposição
ART	Artigo

5.4.1.2. Indexação da *Wikipédia* em um Repositório Local

O *download* automatizado de porções substanciais do conteúdo da *Wikipédia*, obtidos por requisições (acesso ao código HTML por URL) individuais e paralelas a páginas da *Wikipédia* não são permitidas. A política de uso[65] da *Wikipédia* recomenda que, se houver necessidade de acesso maciço aos artigos da enciclopédia, as consultas deverão ser executadas em outro ambiente, por intermédio do *download* de uma cópia da base de dados (*dumps*).

Durante a fase de desenvolvimento, por falta de conhecimentos da política, a primeira versão da arquitetura usou a priorização de uma fila para acessar as páginas da *Wikipédia* de maneira direta. Após "evolução" do protótipo, realizou-se o acesso de maneira paralela (20 acessos simultâneos) à medida que novos conteúdos eram necessários. Isso gerou um grande número de requisições e fez com que a *Wikipédia* aplicasse a penalidade de bloqueio preventivo ao seu conteúdo por 48 horas para o IP requisitante. Com base na política de uso, todo o acesso ao conteúdo da *Wikipédia* passou a ser realizado em um repositório local, de acordo com os seguintes passos.

[65] http://en.wikipedia.org/wiki/Wikipedia:Bot_policy

- Passo 1 – *Download* [66] do arquivo (*dump*) "ptwiki-20111112-pages-articles.xml" (versão 12-11-2011) que contém o conteúdo de toda a *Wikipédia* em português. É valido informar que o processo de verificação e atualização por novas versões do dump depende de intervenção humana.
- Passo 2 – Instanciação do modelo *Wiki* a partir do arquivo do passo anterior. A biblioteca *WikiModel*[67], cria o modelo para os documentos *Wiki* e foi usada para a limpeza dos dados dos artigos que compõem o *dump* da *Wikipédia*. Aplicou-se o filtro *PlainTextConverter* no modelo para a aquisição do conteúdo textual puro (ou seja, isento de itens de formatação) para cada artigo da *Wikipédia*.
- Passo 3 – Geração do Índice no Lucene. Para cada artigo é gerado um *Document* Lucene distinto, composto por quatro *Fields* indexados, usados para recuperação e filtragem (identificador, título, título stemizado e idioma) e um campo (*Field*) não indexado que armazena o conteúdo dos artigos livre de formatação (*content*). A indexação da *Wikipédia* visa ganhos de desempenho em termos de acesso ao conteúdo indexado.

5.4.2. Processamento da Consulta e Extração dos Termos

A Figura 27 ilustra os componentes utilizados na arquitetura física do "Processamento da Consulta", responsável pela extração de termos e geração da requisição de busca que será utilizada pelo protótipo para a expansão das consultas do usuário. A expressão de busca é capturada pelo componente "Seleção de Consulta", que utiliza a biblioteca *RestFB* para comunicação com o *Facebook*. Apenas as mensagens marcadas como expressões de busca são selecionadas. Os termos informados na

[66] http://dumps.wikimedia.org/ptwiki/
[67] http://code.google.com/p/wikimodel/

expressão de busca são submetidos ao processo de "Tokenização, Remoção de *Stopwords* e *Stemming*" e apresenta tratamento especial para os bigramas e trigramas identificados em dados abertos, que serão unidos e tratados como um único *token*. Por exemplo, se o trigrama "rio de janeiro" é informado na expressão de busca e houve casamento deste trigrama em dados abertos, deve-se considerar o *token* "rio_de_jan".

O componente "Seleção do Contexto Enriquecido" utiliza a biblioteca *Semantic Vectors* para associar os termos presentes na consulta do usuário a um conjunto de documentos do contexto enriquecido (ou seja, provenientes do processamento descrito na seção 5.4.1) e seus respectivos rótulos. Para esta pesquisa, consideraram-se os seis (6) documentos com maior *score* entre termos da consulta e documentos do índice, obtido pelo algoritmo de LSA do pacote *Semantic Vectors*. Utilizou-se o rótulo dos documentos para a geração das sugestões do "Grupo A". Estes rótulos foram transformados, de acordo com uma regra de formação. Os rótulos com até três (3) palavras, devem aparecer entre aspas (busca com a frase exata). Utilizou-se a busca por frase exata até três termos, pois este modo de busca apresentou bons resultados em (JOHNSON *et* al., 2006). O mesmo é válido para as palavras dispostas nos parênteses de desambiguação (quando existir). Por exemplo, o artigo "C (linguagem de programação)" usa a expressão entre parênteses (desambiguação) para diferenciar este artigo dos demais, como o artigo que trata da letra "C" do alfabeto. Neste exemplo, os termos dispostos entre parênteses são adicionados à expressão de busca entre aspas, para que a busca ocorra pela frase exata. Caso o rótulo seja composto por quatro (4) ou mais palavras, todas as palavras serão consideradas na expressão de busca de maneira individualizada, ou seja, sem restrição de ordem para os termos.

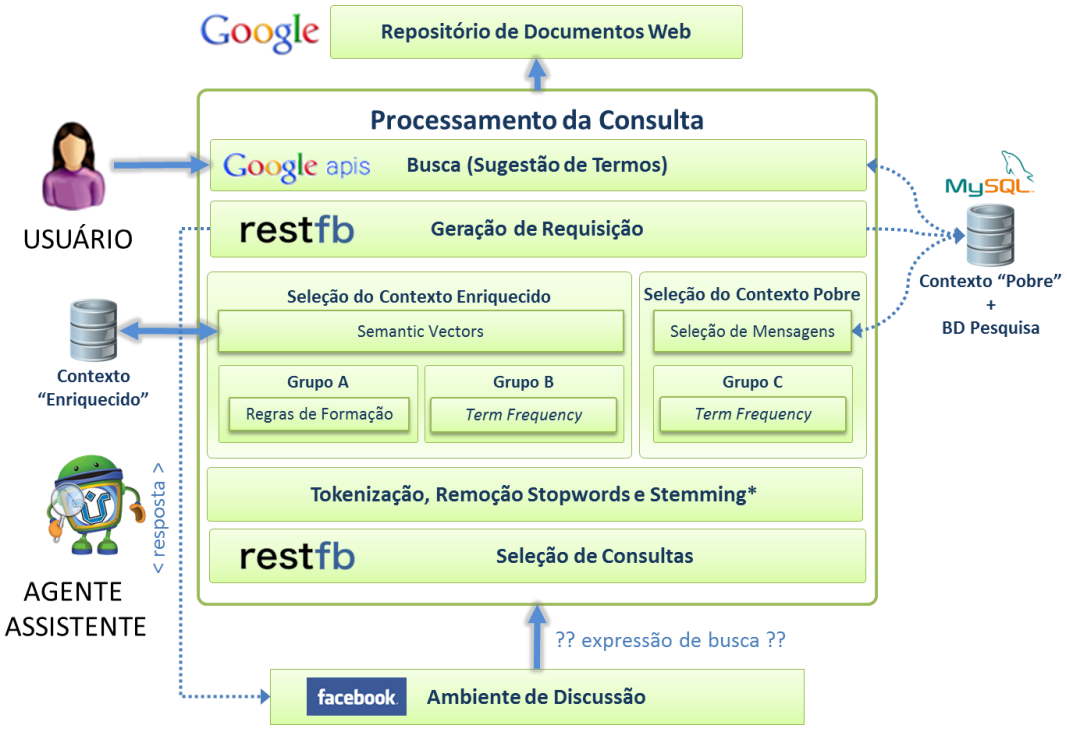


Figura 27 – Componentes da Arquitetura Física da Busca

Para o "Grupo B" considerou-se as palavras mais frequentes contidas nos documentos retornados pelo "Grupo A". A frequência dos termos foi obtida com a aplicação do logaritmo na fórmula de frequência de termos (MANNING *et* al., 2008). Cada artigo pode ser visto como um documento. O cálculo da frequência (*term frequency*) aliado à aplicação do logaritmo visa reduzir a influência negativa de possíveis artigos compostos por muitos termos (normalização). Caso o termo mais frequente fosse um bigrama ou trigrama, o mesmo seria tratado entre aspas.

Já o componente "Seleção do Contexto Pobre" obtém um instantâneo dos termos mais relevantes nas mensagens no momento da busca, ou seja, utilizaram-se apenas os termos mais frequentes das discussões para a geração do "Grupo C". A frequência dos termos foi obtida com a aplicação do logaritmo na fórmula de frequência de termos (MANNING et al., 2008). Cada mensagem pode ser vista como um documento. O cálculo da frequência (*term frequency*) aliada à aplicação do logaritmo visa reduzir a

influência negativa de possíveis mensagens compostas por muitos termos (normalização) frente a mensagens curtas.

O objetivo do componente "Requisição" é unificar todos os termos sugeridos nos grupos A, B e C e grava-los em banco de dados e, em seguida, informar ao usuário solicitante, que a busca está disponível.

Por fim, o componente "Busca" é usado pelo usuário para acessar os documentos *Web* e realizar as avaliações de relevância. Este representa a parte gráfica do protótipo, denominada CCSA (*Collaborative Context Search Agent*). Utilizou-se o padrão MVC (*Model View Controller*), composto por JSP (*Java Server Pages*) e HTML na interface, controlador Servlet que trata a entrada de dados e o modelo do domínio descrito por Entidades JPA (*Java Persistence Architecture*). A recuperação de informações na *Web* usou a *API* Google Search configurada para páginas em português do Brasil, seguindo a política de uso fornecida pela Google Inc. A biblioteca *Google API para Java*[68] fornece acesso a diversos serviços do Google, como o AdSense, Calendar, Analytics, Freebase, Groups, Google+, *Orkut*, Translate etc. Usou-se o *CustomSearch API*[69] para obter resultados de busca personalizada do Google. A administração é feita pelo Google Console *API*[70].

A Figura 28 mostra o diagrama de atividades que trata o processamento da consulta do usuário.

[68] http://code.google.com/p/google-api-java-client/
[69] https://developers.google.com/custom-search/
[70] https://code.google.com/apis/console

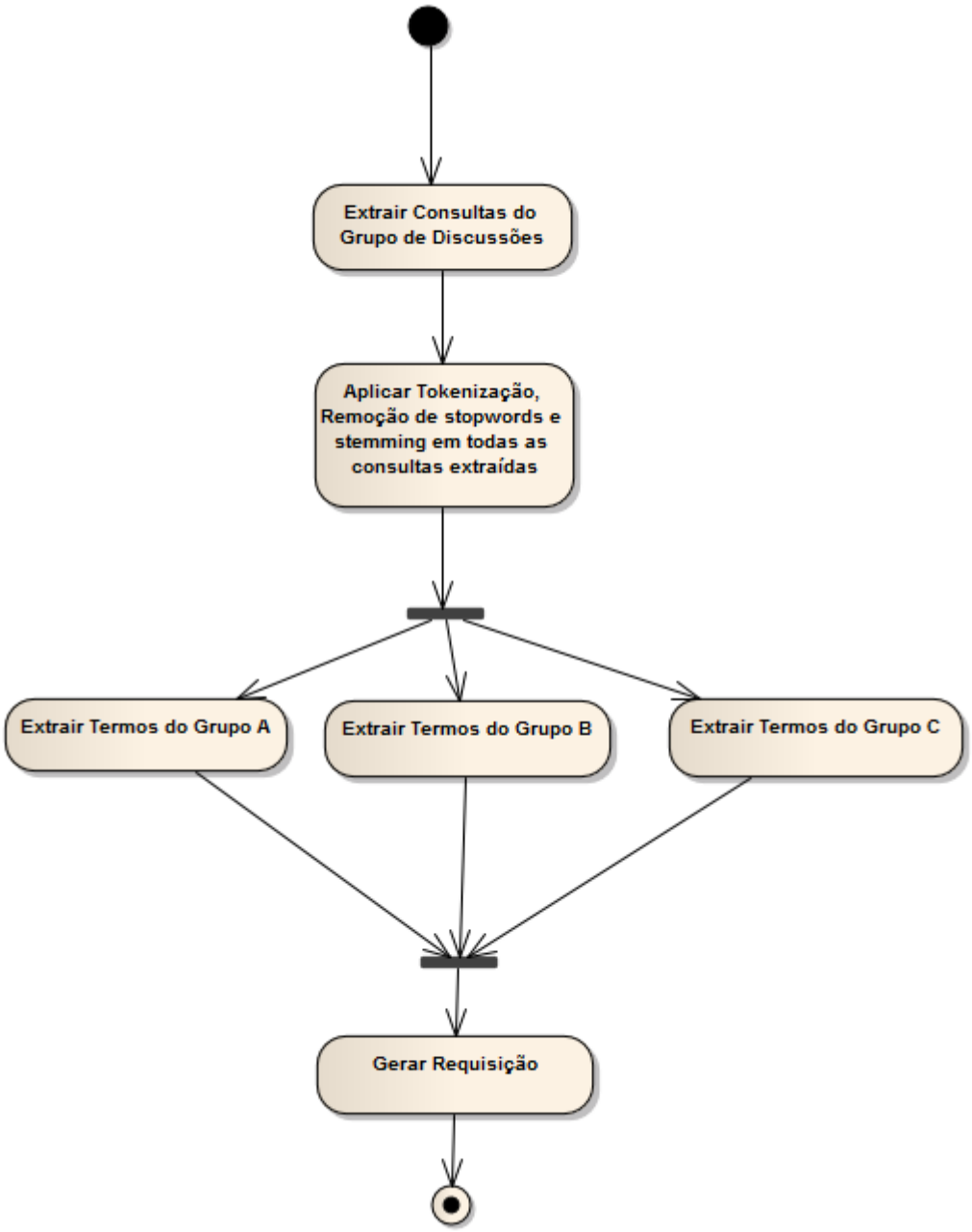


Figura 28 – Diagrama de Atividades para o Processamento da Consulta

5.4.3. Agente de Interface e Modelagem Orientada a Agentes

Este trabalho utiliza os conceitos de agente de interface e modelagem de sistemas orientada a agentes. De maneira geral, considera-se que os agentes são programas que atuam em nome de humanos para executar tarefas trabalhosas como coleta e processamento de informações. O papel do agente no contexto desta dissertação é obter mensagens de um grupo de maneira incremental, com atualização contínua e automática

e trabalhá-las para que se tornem úteis aos seus usuários. Nesse âmbito, o presente trabalho usa a metáfora de "assistente pessoal" (MAES, 1995) e o conceito desenvolvido por Bradshaw (1997), o qual afirma que os agentes devem conhecer seu contexto, ou seja, características e peculiaridades do ambiente, para que possam construir um modelo que irá ajudar na realização de tarefas complexas.

Conforme ilustra a Figura 29, adaptada de (MAES, 1994), os agentes não são o único meio de acesso para o usuário ao computador. Os usuários (neste trabalho alunos) podem interagir diretamente tanto com a aplicação (nesse caso um grupo de rede social), quanto um agente assistente.

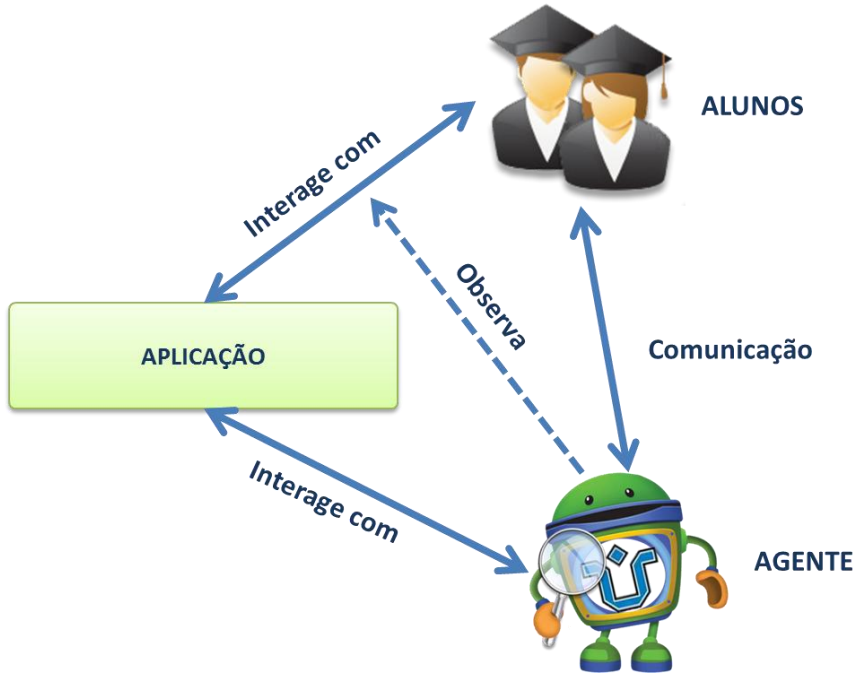


Figura 29 – Interface entre Usuário, Agente e Sistema

Já a proposta de arquitetura multiagente, sob a perspectiva da engenharia de *software* orientada a agentes (HENDERSON-SELLERS E GIORGINI, 2005), visa desenvolver os elementos ativos da arquitetura (WOOLDRIDGE, 2002). O uso da abordagem baseada em agentesé justificado pela necessidade de se manter o contexto atualizado em um ambiente dinâmico e por apresentar vantagens relacionadas à distribuição de tarefas, reutilização e expansão de módulos (agentes de *software* e seus

comportamentos) da arquitetura, como, por exemplo, incluir suporte a outras redes sociais. Os comportamentos utilizados são apresentados no "Apêndice III – Comportamentos do Jade Utilizados".

A arquitetura baseada em agentes é composta exclusivamente por agentes reativos simples, de acordo com as fases de análise e *design* da metodologia MaSE (DELOACH, 2001). O comportamento de cada agente é codificado em um conjunto limitado de regras condição-ação. A vantagem inerente deste tipo de abordagem é a simplicidade e eficiência, embora não se adapte a condições diferentes daquelas previstas pelo projetista. Cada agente proposto possui uma função específica e, por intermédio da comunicação, pode atingir o objetivo global que é sugerir termos para a expansão de consultas. A comunicação é feita pela troca de mensagens ACL sob a ontologia *enrichment-ontology*. Todos os agentes conhecem esta ontologia, que traz informações elementares como a identificação do ambiente, um conjunto de mensagens associadas aos seus criadores e a data/hora de criação.

A arquitetura inclui seis classes de agente, *LexicalAgent*, *EnrichmentAgent*, *SearchAgent*, *ContextManagetAgent*, *InterfaceMonitorAgent e FactoryAgent*. Todos os agentes estendem a classe do pacote "*jade.core.Agent"* e associam-se a, pelo menos, um comportamento, definido no método de inicialização (*setup()*). O *FactoryAgent* (Figura 30) é um agente único na arquitetura, responsável por criar, atualizar e destruir todos os demais agentes, por intermédio da comunicação com o agente AMS. Este agente possui o comportamento *FactoryBehaviour:TickerBehaviour* responsável por verificar, periodicamente, alterações nas configurações dos agentes sob sua responsabilidade na plataforma.

Os agentes *LexicalAgent*, *EnrichmentAgent*, *SearchAgent* são agentes de serviço, instanciados de acordo com a demanda de serviço por seus comportamentos. Estes

agentes devem se registrar no DF (*Directory Facilitator*) da plataforma *Jade* para compor o POOL de agentes responsáveis pelo atendimento de diversos ambientes (diversos grupos de discussão, por exemplo). Já para os agentes *InterfaceMonitorAgent* e *ContextManagetAgent* é criada uma instância por ambiente para cada agente e se utilizam os agentes de serviço com distribuição uniforme das chamadas, com o objetivo de balancear a carga e aumentar o *throughput* do sistema.

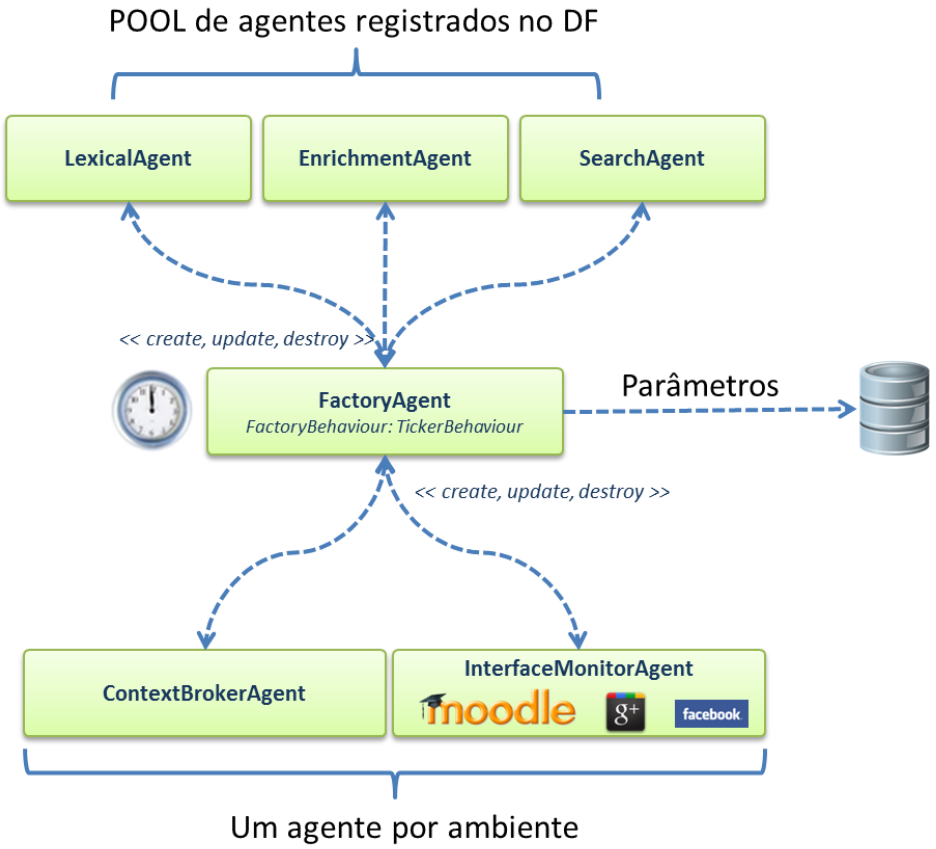


Figura 30 – Criação e Manutenção dos Agentes

A Figura 31 ilustra a arquitetura de colaboração entre os agentes. Nessa imagem é possível identificar, além dos agentes da arquitetura *Jade*, o agente assistente (Agente de interface denominado *UnibotAgent*) e os agentes humanos (Alunos).

O agente *InterfaceMonitorAgent* é um agente responsável por monitorar a interface (*ListenerBerhaviour*) por novas mensagens ou solicitações de consulta dos usuários a partir de grupo de discussões, direcionando-as para que sejam processadas.

Também é responsável por notificar os usuários (*SpeakerBerhaviour*), que sua solicitação de consulta foi processada e que já é possível utilizar o protótipo para a realização de consultas. Este agente conhece o estado atual e o compara ao estado anterior para que somente novos eventos sejam direcionados aos seus respectivos agentes colaboradores. De todos os agentes, este é o único que é instanciado de acordo com a tecnologia do ambiente, como o *Open Graph* adotado pelo *Facebook* e desenvolvido nesta dissertação. Outros ambientes poderiam ser implementados a partir da criação de novos comportamentos para este agente, como *Google+*, *Moodle* etc.

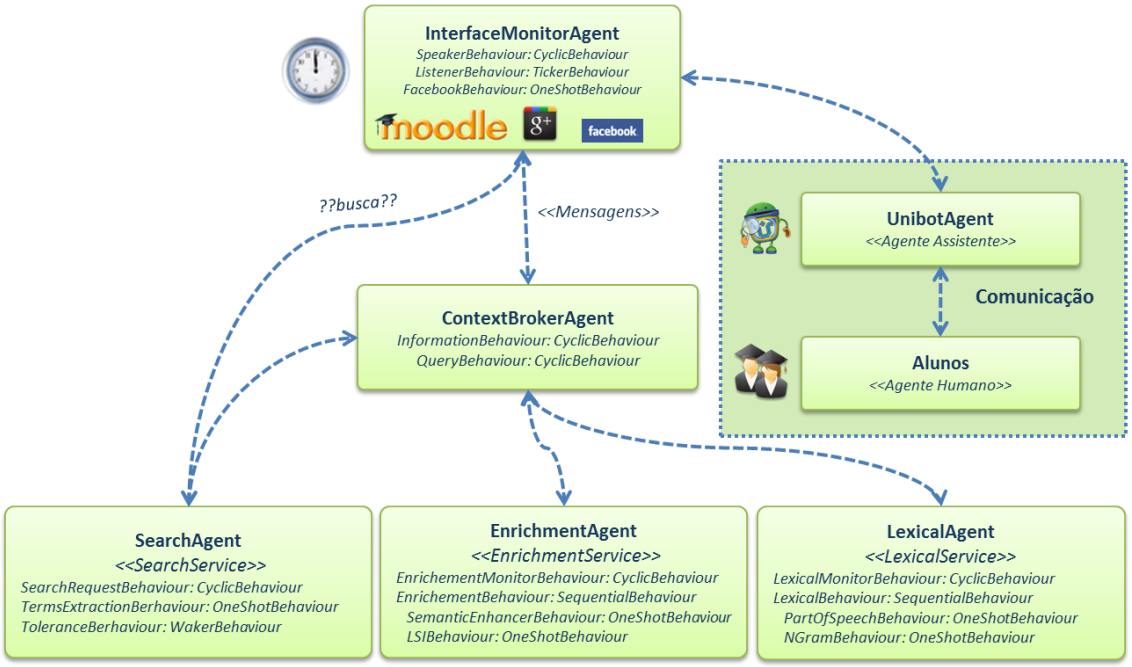


Figura 31 – Arquitetura de Agentes

O *ContextBrokerAgent* é um agente que executa e coordena tarefas relacionadas à construção do contexto. As tarefas de análise léxica das mensagens e enriquecimento em dados abertos são delegadas a outros agentes por intermédio da troca de mensagens. Ele registra as mensagens que foram processadas (conhece o contexto) e sabe quais bigramas e trigramas já foram extraídos. Com isso, para mensagens que contenham bigramas ou trigramas anteriormente encontrados, esta mensagem é tratada como processada e não é enviada ao *EnrichmentAgent*. Este agente acessa informações do

contexto enriquecido e também a base da pesquisa (que inclui mensagens puras, sem o enriquecimento).

SearchAgent é um agente de informação que realiza a recuperação de informação nos contextos enriquecido e não enriquecido para então extrair os termos para expansão (*TermsExtractionBerhaviour*). Recebe uma requisição de busca que deve ser respondida em tempo pré-estabelecido denominado tolerância para atendimento. Para tanto possui um contador de tempo, controlado pelo comportamento *ToleranceBerhaviour: WakerBehaviour*, para aguardar que mensagens dependentes (aquelas geradas antes da busca) sejam atendidas. Se a tolerância esgotar e não houver resposta do *ContextBrokerAgent* sobre a conclusão das mensagens dependentes, o agente processa a extração de termos com o contexto corrente. Este agente acessa informações do contexto enriquecido e também a base da pesquisa, grava a informação de requisição na base da pesquisa e solicita que o agente *InterfaceMonitorAgent* comunique ao solicitante uma mensagem de resposta, que é personificada pelo *UnibotAgent*.

O agente *LexicalAgent* oferece o serviço para tratamento léxico sobre as mensagens e busca por padrões de bigramas e trigramas no texto.

O agente *EnrichmentAgent* é um agente de serviço que busca os bigramas e trigramas encontrados pelo *LexicalAgent* e atualiza o contexto enriquecido com dados oriundos do comportamento de enriquecimento.

5.5. Protótipo *Collaborative Context Search Agent*

O segundo protótipo de aplicação desenvolvido, chamado “CCS Agent” (*Collaborative Context Search Agent),* contemplou os requisitos definidos após o primeiro estudo de caso. Nesse protótipo, a busca por documentos *Web* é feita, a qualquer momento, com o auxílio de um agente assistente, solicitado a partir da própria

interface de discussão. Pensou-se em uma experiência de busca capaz de entregar ao usuário a resposta para a sua necessidade de informação, paralelamente às atividades colaborativas, de maneira a auxiliar a construção do conhecimento desejado sobre o tema no momento que ele ocorre. Essa necessidade de informação por novos conhecimentos deve ser indicada na forma de uma expressão de busca e direcionada ao agente de interface, com o uso de sintaxe específica. A sintaxe deve contemplar a necessidade de informação disposta entre sinais duplos de interrogação, conforme exemplificado no fluxo "C" da Figura 33.

O sistema deve ser capaz de processar a solicitação do usuário e devolver uma mensagem com o *link* para o protótipo, que sugere termos a serem incluídos na consulta e usa o Google como provedor de documentos *Web*. Assim ele evita acessar outro ambiente para realizar a consulta, enquanto aguarda a mensagem de resposta no próprio grupo. A intenção é manter o usuário no grupo enquanto o sistema baseado em agentes processa sua requisição, visto que a resposta do sistema não é imediata.

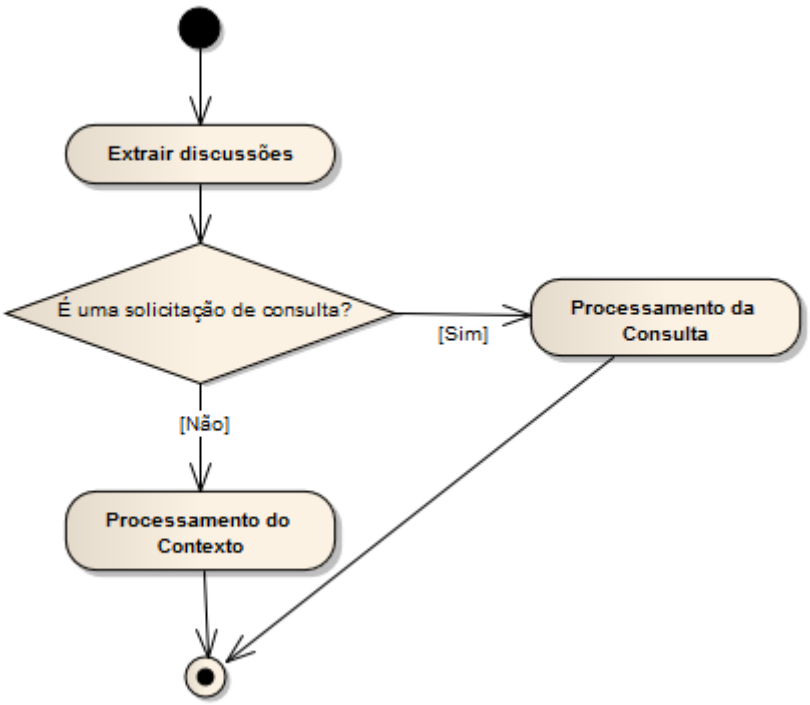


Figura 32 – Diagrama de Atividades para Tratamento das Mensagens

A Figura 33 ilustra a visão geral da proposta de solução, para facilitar o entendimento da arquitetura. Duas macrofuncionalidades, desenvolvidas com a tecnologia de agentes de *software*, podem ser identificadas na figura: processamento do contexto e processamento da consulta do usuário. O direcionamento das mensagens a seis respectivos componentes é feito de acordo com o diagrama de atividades da Figura 32.

Figura 33 – Visão Geral do Enfoque de Solução

O processamento do contexto é alimentado por mensagens do grupo (setas A e B), ou seja, a colaboração entre os usuários irá gerar conteúdo que servirá de insumos para a geração do contexto. Já o processamento da consulta inicia-se pela identificação de uma solicitação de consulta (C), processamento da requisição, que extrai os termos sugeridos para expansão e formula uma mensagem de resposta com o *link* para o protótipo associado à requisição gerada (D). O usuário deve clicar nesse *link* (E) para acessar a

interface do protótipo de busca, escolher os termos para expandir a consulta e avaliar os resultados da busca, clicando sobre as estrelas (F). O desenvolvimento do protótipo de busca baseou-se no padrão MVC[71] (*Model-view-controller*) e empregou as tecnologias JSP (*Java Server Pages*), Servlet e JPA (*Java Persistence Architecture*) com mapeamento das Entidades descritas na seção 5.5.4 em uma base relacional MySQL. Para a gravação das avaliações, adotou-se programação em *Javascript* e Ajax (usou-se a biblioteca jQuery[72]), para a submissão assíncrona das avaliações e animação ao se clicar sobre as estrelas, garantindo o *feedback* para o usuário que sua avaliação foi gravada. Isso objetivou reduzir o número de requisições ao servidor e manter o processo de avaliação mais ágil e fluido, visto que se evitou a resubmissão de toda a interface ao cliente a cada avaliação.

A resposta é enviada pelo agente como uma mensagem, contendo o *link* para o protótipo (Figura 35), que é personalizado para sua necessidade de informação. Cada nova requisição gera três conjuntos de termos, extraídos de maneira distinta e vinculados à expressão de busca informada pelo usuário. A busca é realizada pela combinação dos termos selecionados e gera um conjunto de documentos como resultado. O usuário deve marcar todos os termos que julgar necessário para serem utilizados na expansão de consulta. Termos dos grupos “A”, “B” e “C” podem ser combinados. Para cada combinação inédita de termos, duas consultas são enviadas ao *Google Search API*. A primeira contém os termos originais da consulta e a segunda, além dos termos da consulta original, os termos escolhidos para expansão da consulta. A posição, o *link* e a descrição de cada documento retornado são armazenados como um instantâneo no banco de dados (*snapshot*). Caso o usuário troque a combinação de termos e volte a uma combinação já realizada, será exibido o mesmo conjunto de documentos e suas

[71] Padrão que tem por objetivo separar o sistema em camadas de negócio e apresentação.
[72] http://jquery.com/

respectivas avaliações (estrelas), obtidos exclusivamente do banco de dados, sem acesso ao Google. Isso é importante, visto que os resultados da busca retornados pelo Google poderiam variar com o tempo e o instantâneo do momento da execução da consulta permite que os usuários realizem avaliações parciais e as completem a posteriori. A última combinação de termos é a consulta válida do usuário e deve ser considerada para fins de avaliação caso todos os documentos tenham sido pontuados.

As sugestões de termos (Figura 34), que são mostradas para o usuário na interface do protótipo de busca e avaliação, são organizadas em três (3) grupos de até seis (6) itens (possíveis repetições entre grupos são eliminadas da interface).

Figura 34 – Grupos de Termos Sugeridos

Os termos que são exibidos para o usuário são persistidos no banco de dados (entidades requisicao e requisicao_detalhes).

5.5.1. Avaliação de Relevância

A avaliação de relevância consiste na marcação de estrelas para cada *link* do conjunto de resultados (Figura 35). O padrão para cada nova consulta é de nenhuma estrela, indicado pela cor cinza, sem preenchimento. A avaliação de todos os resultados da busca é obrigatória e pode variar entre seis (6) e doze (12) resultados. A variação no

número de resultados consolidados (exibidos na interface) está relacionada à possível sobreposição de resultados entre a consulta original e a consulta expandida. Não são apresentados *links* redundantes. Com isso, na ocasião de todos os *links* coincidirem, somente seis (6) *links* são apresentados e caso nenhum coincida serão apresentados doze (12) *links*. Isso procura evitar redundâncias e, consequentemente, reduz a necessidade de reanálise e retrabalho do usuário (evita a perda de tempo), além de manter a consistência da avaliação. Caso um *link* ocorra nos resultados da consulta original e expandida, este será armazenado com a mesma avaliação para ambas e apresentado somente uma vez ao usuário.

Este protótipo armazena dados importantes para a avaliação dessa pesquisa, que envolve a comparação de relevância entre os resultados com e sem o uso das expansões de consulta. Todas as avaliações geradas estão associadas aos seus respectivos avaliadores. Para a identificação dos usuários, usou-se o mecanismo de autenticação da própria rede social.

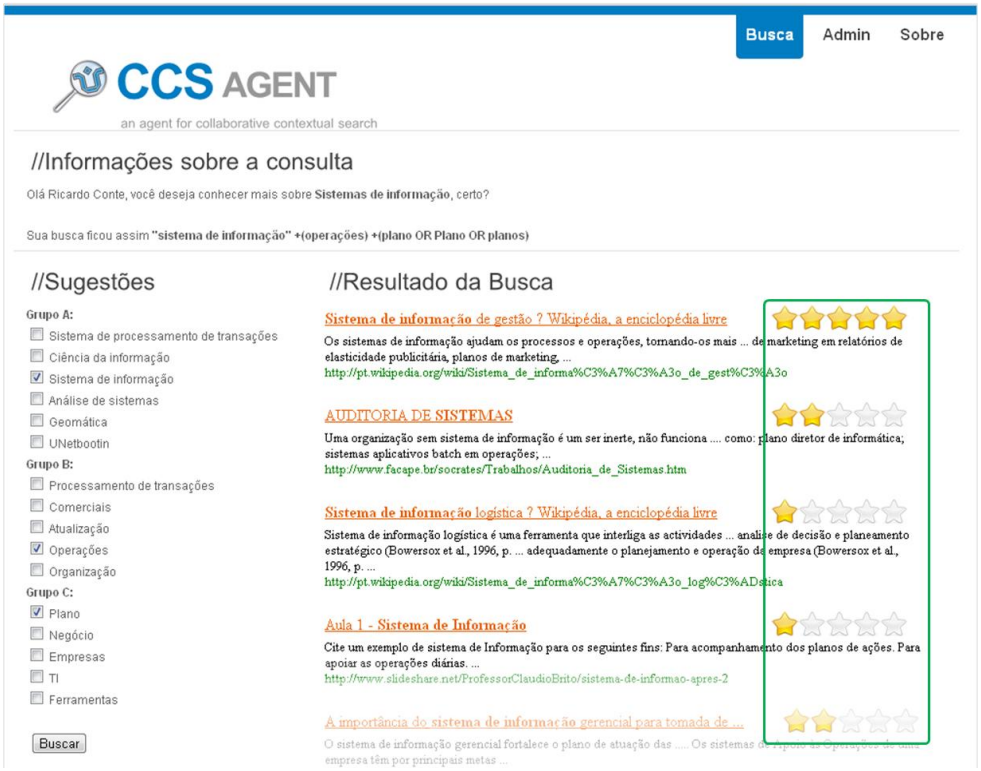


Figura 35 - Avaliação de relevância dos resultados de busca.

5.5.2. Administração de Grupos

A administração de grupos (Figura 36) visa adicionar novos ambientes a serem assistidos e atendidos pela proposta. Cada novo ambiente é associado a um contexto isolado, ou seja, dois ou mais grupos de discussão podem coexistir, cada qual com seu contexto. Para este protótipo, desenvolveu-se o suporte a grupos do *Facebook*. Entretanto, é possível e desejável que este protótipo possa evoluir para prover suporte a outros ambientes, como outras redes sociais que não o *Facebook*, fóruns de discussão, plataformas de aprendizagem etc.

Dentre os parâmetros para a definição de um novo ambiente estão: (i) nome do ambiente a ser monitorado, (ii) identificador do grupo de discussões, (iii) intervalo de verificação de alterações no ambiente, (iv) idioma das discussões e (v) outros parâmetros específicos do ambiente em questão como, no caso do *Facebook*, as chaves de identificação da aplicação, *token* e *secret*. Para cada novo ambiente deve ser criado um novo agente *InterfaceMonitorAgent* para monitorar os eventos da interface.

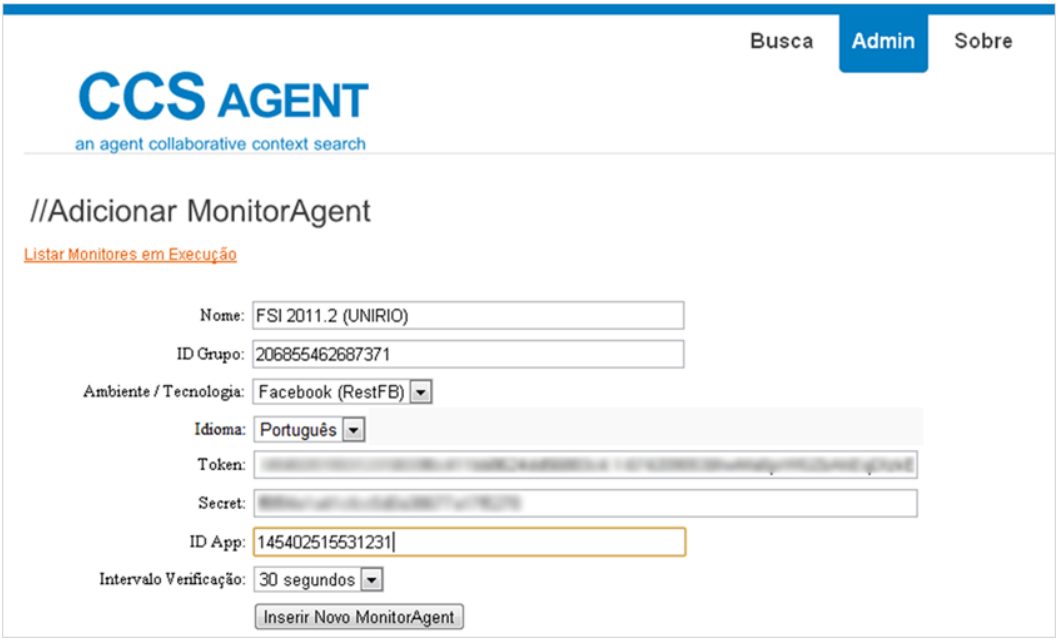


Figura 36 – Visão Administrativa para a Configuração do Protótipo

5.5.3. Recuperação a Falha

Todas as informações processadas pela arquitetura são persistidas em banco de dados. O contexto de execução é recuperado em caso de falha. As informações incluem: mensagens, mensagens anotadas linguisticamente, usuários, requisições de consulta, índice de documentos, vetores semânticos, histórico das buscas e avaliações realizadas no protótipo.

5.5.4. Modelagem de Dados

O modelo de dados (Figura 37) é composto por 10 tabelas. A função deste modelo é permitir a rastreabilidade cronológica entre todos os eventos da interação usuário-computador, tanto no ambiente de discussões, quanto no protótipo de busca.

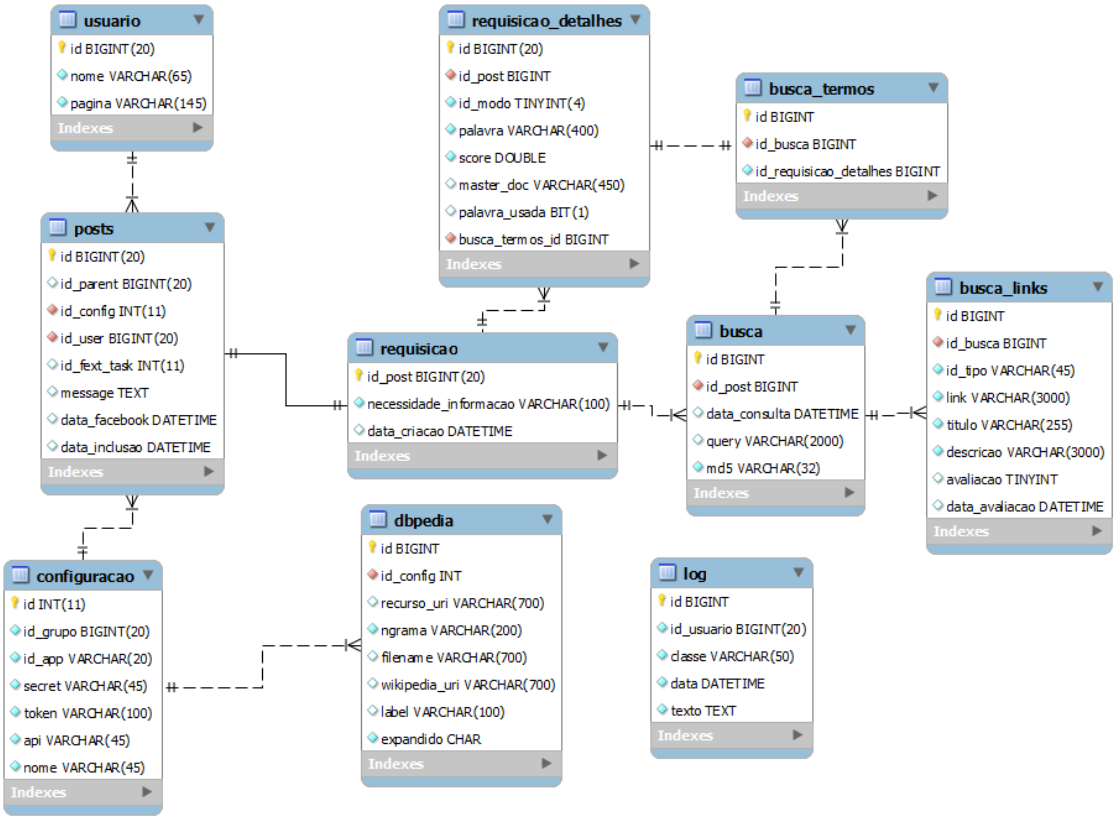


Figura 37 – Modelo de Dados da Arquitetura

- A Tabela “usuario" – armazena informações dos usuários e relaciona-se com a tabela de “posts”. Isso faz com que as mensagens dos usuários possam ser identificadas, como autor e sua respectiva página no *Facebook*.
- Tabela “posts” – conteúdo textual de todas as mensagens (incluindo data de criação e data de extração) de um grupo. Esta tabela relaciona-se com as tabelas “configuração”, para identificação do grupo/ambiente de discussão e “requisição”, para identificação das mensagens de busca (sinais duplos de interrogação). Também é armazenado o identificador de processamento de linguagem natural obtido pelo serviço F-EXT-WS.
- Tabela “configuração” – informações para identificação do contexto. Este protótipo pode atender a múltiplos contextos, atualmente limitados a grupos do sítio de rede social *Facebook*. Armazena informações importantes para a comunicação com a *API* do *Facebook* como identificações da *app*, do grupo e de acesso.
- Tabela “dbpedia” – informações sobre associação entre n-gramas e instâncias da *DBpedia*, relacionadas a um ambiente (tabela “configuração”).
- Tabela “requisicao” – armazena o resultado de uma solicitação de consulta e a data/hora da disponibilização. Relaciona-se com as tabelas “requisição_detalhes”, que detém os termos sugeridos para uma necessidade de informação em um momento no temo (*snapshot*) e “busca”, que contempla todas as buscas a *Web* relacionadas a esta requisição.
- Tabela “requisicao_detalhes” – termos de sugestão associados à requisição e a qual grupo pertencem (A, B ou C).

- Tabela “busca” – todas as buscas realizadas no protótipo, ou seja, todas as vezes que um usuário combinou termos sugeridos e clicou em buscar. Relaciona-se com a tabela “busca_termos”.
- Tabela “busca_termos” – os termos escolhidos pelo usuário para uma busca.
- Tabela “busca_*links*” – representa o conjunto de documentos retornados por uma busca ao Google pelo protótipo que é persistido em banco de dados. Duas buscas são enviadas ao Google a cada combinação inédita de termos e são diferenciadas pelo atributo tipo de cosulta (1=busca_original e 2=busca_expandida). Caso o usuário escolha uma combinação de termos já escolhida anteriormente (verificado nas tabelas busca_termos e requisicao_detalhes), o protótipo irá recuperar o conjunto de resultados desta própria tabela e suas respectivas avaliações, mesmo que parciais. Esta é a tabela mais importante para fins de avaliação, pois armazena a nota atribuída pelo usuário em seu julgamento de relevância.

Capítulo 6 – Segundo Estudo de Caso

Neste capítulo será descrito o segundo estudo de caso, a caracterização dos participantes e seus principais resultados.

6.1. Metodologia

Um estudo de caso único foi conduzido em um ambiente acadêmico (UNIRIO) e teve a participação humana na utilização de um protótipo. Este estudo de caso usa as mesmas métricas de avaliação do primeiro estudo de caso e compara a relevância da busca original (sem expansão) com a busca expandida. Similarmente ao primeiro estudo de caso, os alunos foram instruídos a participar de uma aula baseada em discussão, realizada no *Facebook*. Usou-se a mediação do professor para fomentar as discussões e acompanhar a sua aderência ao tema proposto.

Um roteiro foi elaborado para o estudo de caso e disponibilizado aos alunos. Além do roteiro, o pesquisador aproveitou a reunião dos alunos na aula para explicar a dinâmica das atividades e o objetivo do estudo. O roteiro definiu os objetivos desta pesquisa, bem como o tema da discussão. Os alunos receberam instruções de ingresso no grupo e a maneira como o *Facebook* seria utilizado no estudo. Eles foram instruídos a interagir livremente entre si, a fim de adquirir e trocar conhecimentos acerca do tema proposto. O Apêndice I mostra o roteiro que foi disponibilizado aos alunos para a participação no segundo estudo de caso.

6.2. Dados Coletados

Assim como no primeiro estudo de caso, a avaliação e os resultados foram trabalhados a partir da coleta dos seguintes dados: histórico de mensagens no grupo, *log* do sistema (evolução do contexto no tempo), avaliação das buscas e questionários. Todos os participantes desta pesquisa tiveram suas identidades mantidas em sigilo. Todos os nomes identificados nas figuras são pseudônimos.

O questionário foi utilizado com a finalidade de ser mais um instrumento de coleta de dados e deveria ser preenchido ao final da dinâmica, com o objetivo de adquirir mais evidências que auxiliem a confirmar ou refutar a hipótese. As perguntas foram organizadas em duas partes, com objetivos distintos. A primeira parte do questionário objetivou traçar o perfil dos participantes e a segunda parte avaliar qualitativamente o protótipo (*Collaborative Context Search Agent – CCSA*) e entender, dentre outras coisas, qual grupo de termos foi mais útil à busca. Os resultados foram analisados qualitativamente e comparados aos resultados quantitativos extraídos da avaliação e do *log*.

Novamente, realizaram-se perguntas fechadas de múltipla escolha e perguntas abertas. Usou-se vocabulário simples, sem termos técnicos, com o objetivo de evitar confusão. Preferiram-se perguntas no estilo de múltipla escolha pela facilidade de codificação, tabulação e comparação dos resultados, o que permitiu a comparação das opiniões de maneira objetiva.

Para a maior parte das perguntas de múltipla escolha, afirmações foram feitas acerca do protótipo e as alternativas seguiram a escala de LIKERT (1932). Uma alternativa dicotômica usou o diferencial semântico (melhorou – piorou) para medir a qualidade dos termos sugeridos no tempo (OSGOOD et al., 1957).

As perguntas abertas objetivaram aprofundar a visão e opinião dos alunos, que puderam aproveitar a liberdade para expressar seus sentimentos, comentários e sugestões sobre o protótipo. O Apêndice II contém a íntegra do questionário aplicado no segundo estudo de caso.

6.3. Planejamento do Segundo Estudo de Caso

O segundo estudo de caso foi planejado para avaliar a execução da proposta em um ambiente de aprendizagem. Observações das tarefas e resultados obtidos no primeiro estudo de caso (relacionados à proposta anterior) influenciaram o planejamento deste estudo. A tarefa de avaliação foi simplificada em relação ao primeiro estudo de caso, reduzindo-se a quantidade de avaliações por consulta de cinquenta (50) para doze (12), tendo em vista minimizar os vieses que pudessem ser introduzidos no processo, relacionados ao esforço desprendido à tarefa de avaliação que, nesta nova proposta, é concorrente às discussões. Outro diferencial em relação ao primeiro estudo de caso é que os resultados não são divididos em categorias, como consulta original e consulta expandida, mas sim mesclados em uma única lista. Se houver coincidência de *links* entre os resultados da consulta original e consulta expandida, estes resultados serão unificados, ou seja, as redundâncias serão eliminadas da interface. Ao realizar a avaliação, os *links* coincidentes receberão o mesmo critério (estrela/nota de avaliação), de modo transparente. Mais uma vantagem da unificação é que, na avaliação, o rótulo (“consulta original” ou “consulta expandida”) poderia identificar os resultados da pesquisa, e influenciar a nota dos participantes. Estes cuidados têm o objetivo de não influenciar a avaliação, que deve ser feita de forma imparcial.

Outra questão importante sobre a dinâmica realizada no segundo estudo de caso, é que se esperou um dia para evitar o problema do arranque a frio. A partir daí, permitiu-

se aos alunos a livre interação com o agente assistente, que intermediou a ida ao protótipo de busca. No protótipo, cada aluno pôde executar o número de buscas e combinação de termos que julgasse necessário e avaliaram a relevância de todos os resultados retornados. Nesse âmbito, os documentos na *Web* podem ser vistos como fonte de informação para que os alunos possam entender e participar mais das discussões. Após a realização da dinâmica, os participantes foram convidados a preencher um questionário com questões qualitativas sobre a relevância dos resultados na utilização do protótipo *Collaborative Context Search Agent (CCSA).*

O estudo de caso foi realizado para observar e relacionar a relevância dos resultados retornados pelas duas modalidades de consulta (original e expandida) e a combinação de termos selecionada nos grupos A e B (termos do contexto com enriquecimento) ou C (termos do contexto sem enriquecimento). A relevância de cada documento retornado no protótipo é aferida explicitamente pelos usuários numa escala de cinco (5) estrelas e relaciona-se implicitamente com o(s) grupo(s) do(s) termo(s) selecionado(s). Em outras palavras, espera-se avaliar a relevância dos resultados da consulta original com os resultados da consulta expandida, observando-se o grupo de termos escolhido para a expansão da consulta (ou suas combinações). Foi obrigatória a escolha de pelo menos um termo de qualquer um dos grupos para expansão. Ainda em relação aos três grupos de termos (“Grupo A”, “Grupo B” e “Grupo C”), preferiu-se generalizar os rótulos dos grupos, para que não houvesse identificação e, por conseguinte, influência na escolha dos termos desses grupos. Por exemplo, o nome do “Grupo A” poderia ser “Rótulos Open Data”, que é potencialmente mais pomposo que o “Grupo C”, “Termos das Discussões”.

6.3.1. Perfil dos Participantes do Segundo Estudo de Caso

Participaram do estudo 18 alunos do primeiro período do curso de Bacharelado em Sistemas de Informação, turma BSI 2011.2, matriculados na disciplina "Fundamentos de Sistemas de Informação". Os participantes possuem idade entre 18 e 26 anos, média e mediana de 20 anos. Todos os participantes possuem computador e acesso a internet. Apenas um participante não possuía perfil na rede social *Facebook*, que foi prontamente criado para participar da pesquisa. Do total de participantes, 14 responderam o questionário (1 do sexo feminino e 13 do sexo masculino). Ainda, ao contrário do perfil dos participantes do primeiro estudo de caso, neste, quando perguntados sobre como preferem estudar, a maioria respondeu "Em grupo" (79%).

6.4. Métricas Utilizadas na Avaliação

Neste estudo de caso, foram exploradas as mesmas três métricas de avaliação sensíveis à ponderação de relevância humana: (i) precisão total dos *x* primeiros resultados, (ii) comprimento da busca e (iii) correlação de *ranking*.

A escala de relevância foi a escala de cinco estrelas, na qual uma estrela (sem relevância, insatisfeito) é representada pelo valor "0" e cinco estrelas (relevância máxima, completamente satisfeito) representadas pelo valor "4".

6.4.1. Precisão Total dos *x* Primeiros Resultados (CHIGNELL *et* al., 1999)

A medida de precisão total dos *X* primeiros resultados visa calcular a quantidade de informações relevantes encontradas nos primeiros resultados, onde *Y* é o resultado da multiplicação entre o número total de resultados considerados (*X*) e o valor máximo de relevância atribuído para cada resultado (4). A Tabela 6 ilustra os valores praticados para o segundo estudo de caso.

Tabela 6: Valores para o estudo de caso

Variável	Estudo de Caso 2
Relevância máxima	4
Valor de X	6
Valor de Y	24

Embora o valor originário de X para esta métrica seja 20, houve necessidade de redução da quantidade de resultados retornados de 20 para 6 resultados no segundo estudo. A primeira justificativa para a alteração do número de resultados é a observação do comportamento dos usuários ao realizar buscas na *Web*. Usuários acessam poucas páginas de resultado, frequentemente restringem-se aos cinco primeiros resultados (SPINK e JANSEN, 2004) e (JANSEN *et* al., 1998). Como os resultados da consulta original e expandida poderiam ser sobrepostos, optou-se por incluir um resultado adicional a cada modalidade. A segunda justificativa relaciona-se ao tempo concorrente entre pesquisa e avaliação dos resultados e a participação no grupo de estudo.

É importante ressaltar que os resultados originais (6) e expandidos (6) são exibidos na mesma lista de resultados no segundo estudo de caso. Para fins de consistência, resultados redundantes são exibidos apenas uma vez e recebem a mesma avaliação. Caso os seis primeiros resultados de cada expansão (expressões de busca original e expandida) sejam diferentes, serão apresentados ao usuário doze resultados para avaliação (valor máximo). Caso os primeiros seis primeiros resultados coincidam (caso extremo) para ambas as consultas (original e expandida), apenas seis resultados serão exibidos aos usuários (valor mínimo).

6.4.2. Comprimento da busca (COOPER, 1968)

A segunda métrica utilizada indica o esforço do usuário em examinar resultados de pouca relevância, até encontrar *x* documentos relevantes. Esta pesquisa adotou os parâmetros *x=1* (primeiro documento relevante) e *x=2* (os dois primeiros documentos

consecutivos e relevantes) e relevância igual a quatro ou cinco pontos (estrelas), os mesmos adotados por Tang e Sun (2003).

Vale ressaltar que a proposta de solução difere no número de documentos retornados. Apenas os seis (6) primeiros resultados foram apreciados. Considerando que Tang e Sun (2003) usaram um conjunto composto por 20 resultados, essa redução torna a medida muito mais rigorosa e dá uma ênfase muito maior aos primeiros resultados.

6.4.3. Correlação de *ranking* (SU *et* al., 1998)

Esta última métrica busca estabelecer uma medida de avaliação que correlaciona os *ratings* de relevância (escala de cinco (5) pontos) obtidos pela avaliação qualitativa de seres humanos e o *ranking* atribuído pelo motor de busca para a priorização dos *x* primeiros resultados.

Como não existe acesso às notas reais de classificação atribuídas pelo sistema (Google no caso deste trabalho), utilizou-se a posição do documento no conjunto-resposta para presumir sua pontuação. Quanto maior a proximidade do documento ao topo da lista, maior a sua pontuação. Empregou-se o coeficiente de correlação de Pearson para o cálculo da correlação entre a matriz A, que representa as avaliações dos usuários e a matriz B, que representa a ponderação associada a sua posição no conjunto de resultados da busca. O resultado da correlação entre as variáveis é apresentado no intervalo [-1, +1], onde os resultados próximos de +1 representam uma correlação perfeita positiva, 0 ausência de correlação linear e -1 correlação perfeita negativa entre as variáveis (KENDALL e STUART, 1973).

Para o segundo estudo de caso, avaliou-se seis resultados ($x=6$) para cada tipo de consulta (original e expandida). Duas matrizes foram consideradas para o cálculo de correlação nesse estudo de caso. Na penúltima coluna da Tabela 7 (Correlação'), optou-se pela mesma distribuição adotada no primeiro estudo de caso, porém o valor inicial

para a matriz foi dois, devido ao menor número de resultados. Essa modificação implica em uma redução na correlação de Pearson, afastando-se de um (1) o valor da correlação, uma vez que o valor máximo de avaliação é quatro, não dois. Entretanto, essa consideração não implica em problemas, pois, tanto o tipo de consulta 'Original', quanto o tipo de consulta 'Expandida' são submetidas à mesma matriz e, portanto, podem ser comparadas em termos relativos. A última coluna da Tabela 7, (Correlação''), apresenta uma distribuição uniforme para a correlação, que varia de 4 (pontuação máxima atribuída pelo usuário a um documento) a 0 (pontuação mínima atribuída pelo usuário a um documento), com passo de 0.8.

Tabela 7: Matriz de Correlação Para o Estudo de Caso 2

Sequencial	Resultado	Tipo	Correlação'	Correlação''
01	1	Original	2	4
02	2	Original	2	3,6
03	3	Original	1	2,4
04	4	Original	1	1,6
05	5	Original	0	0,8
06	6	Original	0	0
07	1	Expandida	2	4
08	2	Expandida	2	3,6
09	3	Expandida	1	2,4
10	4	Expandida	1	1,6
11	5	Expandida	0	0,8
12	6	Expandida	0	0

6.5. Avaliação do Segundo Estudo de Caso

No segundo estudo de caso, a avaliação quantitativa consistiu em julgar a relevância dos resultados consolidados de busca obtidos na consulta original e expandida, ou seja, *links* redundantes foram mesclados e exibidos em uma única lista aos usuários. O número de avaliações podia variar entre seis (6) resultados no caso de sobreposição total e doze (12) resultados no caso de não existir sobreposição. Na prática, o número máximo de sobreposição foi quatro (4) e o mínimo foi zero (0). Os alunos

avaliaram 10,85 resultados em média. Em números absolutos, dos 408 *links* avaliados (documentos *Web*), 330 foram *links* únicos e 78 sobreposições. Como destes 78 *links* sobrepostos apenas um foi exibido ao usuário, realizaram-se 39 avaliações explícitas (as outras 39 avaliações foram implícitas).

Para a expansão de consultas interativa, existia à disposição do usuário até seis (6) termos divididos em três (3) grupos (18 termos no máximo). Embora expansão de consultas interativa (*Interactive Query Expansion*) exija esforço adicional do usuário em ler todos os termos e selecionar os mais adequados (JOHO *et* al., 2004) e de requerer maior experiência de busca por parte do usuário (RUTHVEN, 2003), preferiu-se essa maneira à expansão de consultas automática (*Automatic Query Expansion*), pois, além de possuir grande potencial em gerar melhores resultados que a expansão de consulta automática (KANAAN *et* al., 2008), desejou-se identificar o grupo de termos escolhido (avaliação implícita) e a relação de suas combinações com as métricas de avaliação.

Inicialmente, o estudo de caso contou com a participação de 20 alunos, que debateram durante cinco (5) dias, sobre assunto definido pelo professor em um grupo de uma plataforma de rede social. O Apêndice IV – Mensagens Durante a Dinâmica, traz uma parte deste debate. Desses, 18 alunos realizaram pelo menos uma avaliação no protótipo e 14 preencheram o questionário.

O número médio de termos selecionados para expansão foi de 1,7 termos, considerando os três grupos. O número médio de termos informados na consulta original foi de 2,29 termos.

Foram descartadas do estudo avaliações incompletas para uma mesma requisição.

6.5.1. Considerações de Desempenho

Foram 286 mensagens trocadas em cinco (5) dias de experimento. Destas, houve apenas uma ocorrência na modalidade *link*, ou seja, foram consideradas 285 postagens

de mensagens de discussão. A configuração do servidor é a mesma que a relatada na seção 4.5.1.

O *InterfaceMonitorAgent* verificou, a cada 30 segundos, modificações no ambiente. Não houve problemas de limite[73] de acessos aos dados do grupo no Facebook.

Em relação ao tempo de atendimento das requisições, o pior caso foi registrado no primeiro dia de consultas (2 de dezembro de 2011): 19,86 horas (71506 segundos). Este tempo está associado a motivo de força maior e é relatado na seção 6.5.2. Ao considerar o dia com mais solicitações atendidas (3 de dezembro de 2011), o tempo mínimo de atendimento foi de 43 segundos, média de 130 segundos e mediana de 63 segundos.

6.5.2. Análise Quantitativa

A dinâmica teve início no dia 01 de dezembro de 2011 e terminou no dia 05 de dezembro de 2011. O planejamento do estudo reservou o primeiro dia somente para as discussões, para que fosse evitado o problema do arranque a frio. Nesse dia (1 de dezembro de 2011), as requisições feitas ao agente, apesar de instrução ao grupo de proibir consultas no primeiro dia, receberam uma resposta indicando que este ainda não tinha conhecimentos suficientes para realizar a sugestão dos termos (Figura 38).

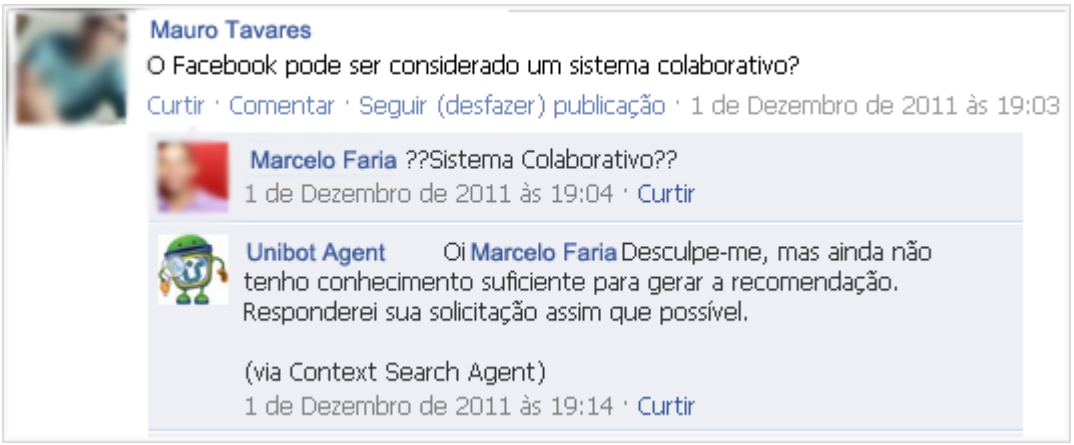


Figura 38 – Mensagem de Desculpas do Agente

[73] http://developers.facebook.com/policy/

No dia seguinte (2 de dezembro de 2011) estava planejada a liberação do ambiente para pesquisas, mas no decorrer desse dia ocorreu um problema de ordem técnica em relação à conexão com a internet no *link* do servidor do experimento. Isso fez com que os alunos ficassem sem respostas as suas solicitações e justifica o baixo número de avaliações para o primeiro dia de consultas. Para os demais dias, é possível notar a queda na participação no decorrer do estudo. A Figura 39 ilustra o número de avaliações consideradas durante os quatro dias que o sistema ficou disponível. Foram consideradas 06 avaliações para o dia 02 (18%), 14 para o dia 3 (41%), 11 para o dia 4 (32%) e 3 para o dia 5 (9%).

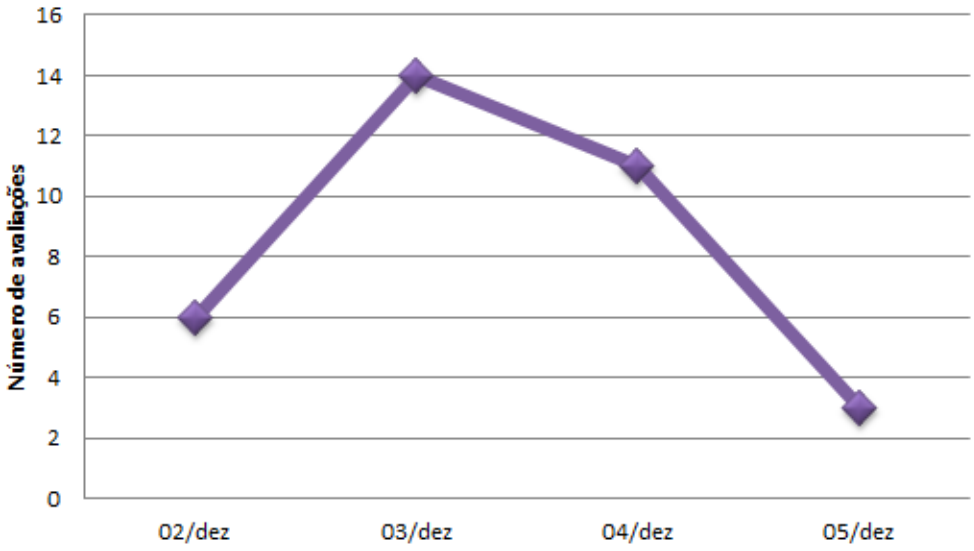


Figura 39 – Avaliações Consideradas por Dia

A Figura 40 expande a Figura 39 e contrapõe o número de avaliações consideradas com o número de mensagens extraídas do grupo em cada dia do estudo. A partir da figura, percebe-se que não existiram avaliações consideradas (visto que as consultas não estavam disponíveis no primeiro dia de pesquisa (01/dez)) e participação modesta na quantidade de mensagens para este dia. No dia seguinte (02/dez) percebe-se baixa quantidade no número de avaliações consideradas, tendo em vista problemas de ordem técnica com o *link* de conexão com a internet no servidor do estudo de caso.

Apesar disso, como o ambiente de discussão (*Facebook*) continuou operável, percebe-se grande interação entre os participantes do estudo no grupo (136 mensagens). O número de consultas deste dia foi atendido no dia seguinte (03/dez).

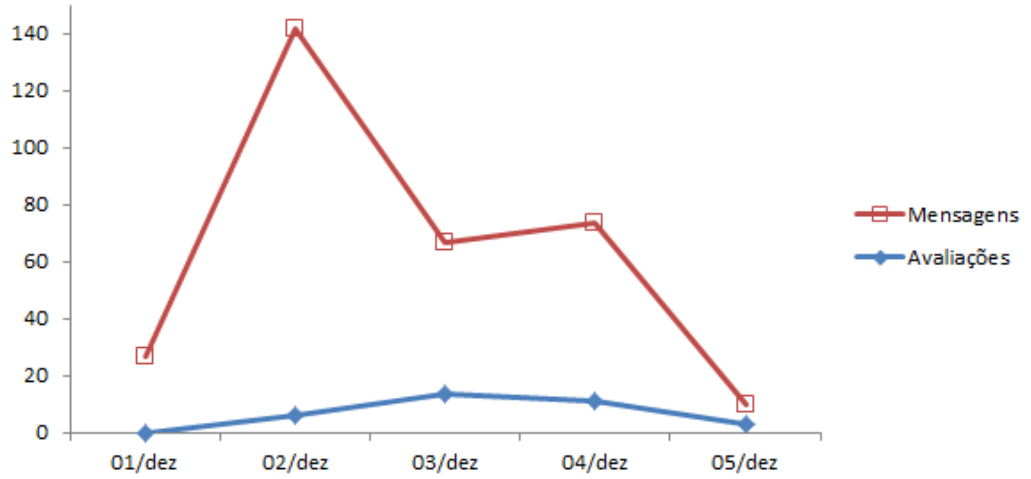


Figura 40 – Avaliações Consideradas por Dia, em Relação às Mensagens

Foram selecionadas 58 sugestões de termos (nos três grupos) para as 34 avaliações consideradas, uma média de 1,7 termos escolhidos por consulta. De todas as avaliações consideradas, a maior parte dos alunos usou os termos sugeridos no "Grupo A" (rótulos dos artigos), frente ao "Grupo B" (termos mais frequentes no conteúdo dos artigos do Grupo A) e "Grupo C" (termos mais frequentes nas discussões), conforme mostra a Figura 41. Ainda, o uso dos termos para o grupo "B" foi 8% superior em relação ao grupo "C".

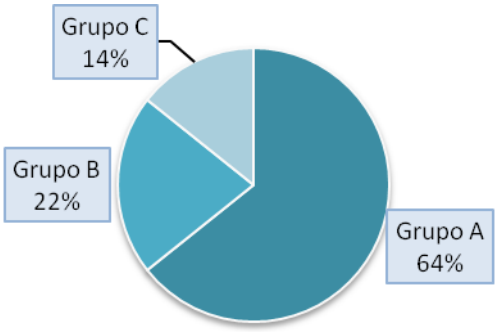


Figura 41 – Grupo de Termos Escolhidos no Protótipo

Esse resultado está de acordo com os dados do questionário que também apontam o Grupo A como aquele que trouxe melhores resultados a consulta (Figura 42). Uma pequena divergência em relação ao protótipo pode ser vista entre os grupos B e C, a qual mostra uma superioridade de 9% para o grupo C.

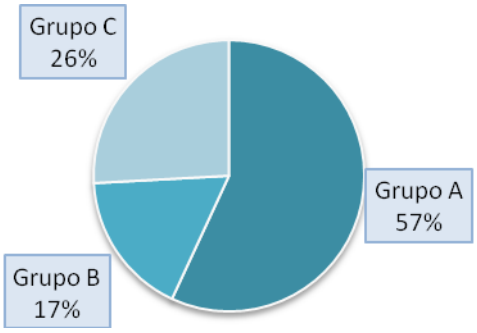


Figura 42 – Grupo de Termos Preferido Segundo o Questionário

Com relação à precisão total, a consulta expandida trouxe melhores resultados que a consulta original em todos os dias, conforme observado na Figura 43. As consultas expandidas apresentaram melhores resultados em 67% dos casos nos dias 2 e 5, 57% no dia 3 e 55% no dia 4. As consultas originais foram melhores que as consultas expandidas em 17% (dia 2), 36% (dia 3), 27% (dia 4) e 33% (dia 5) dos casos. Com isso, a precisão média de pesquisa para a consulta original foi de 29% e de 62% para a consulta expandida.

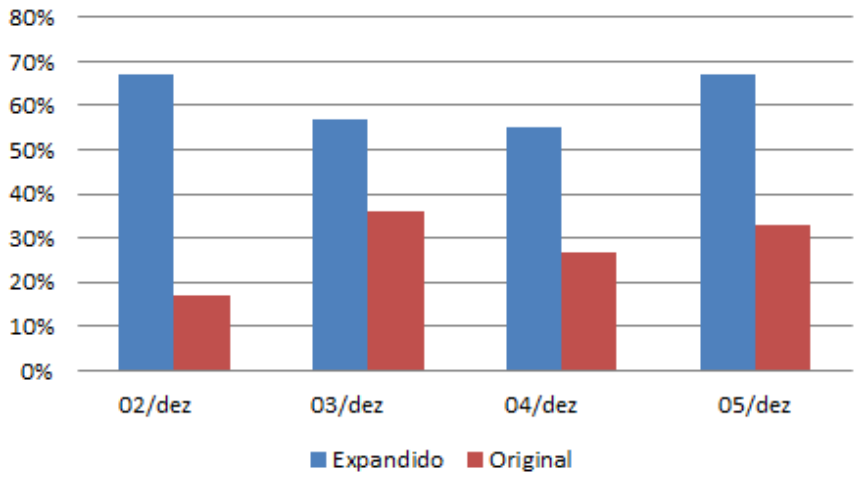


Figura 43 – Precisão Total por Dia de Pesquisa

Ao desmembrar todas as possibilidades de combinação entre os grupos A, B e C, surgem sete (7) possibilidades. A combinação e a quantidade de vezes que a mesma ocorreu no estudo são descritas na Tabela 8.

Tabela 8: Combinação dos Grupos de Termos

Combinações	Quantidade
Somente A	18
Somente B	2
Somente C	4
A e B	3
A e C	2
B e C	3
A, B e C	2

A precisão total consolidada para os 5 (cinco) dias de estudo, observada sobre a ótica das combinações é descrita na Figura 44. A consulta original foi melhor nos grupos {A, C} e {B, C}. Já a consulta expandida foi melhor nos Grupos {A} e {A, B}. Os grupos {B}, {C} e {A, B e C} apresentaram empate.

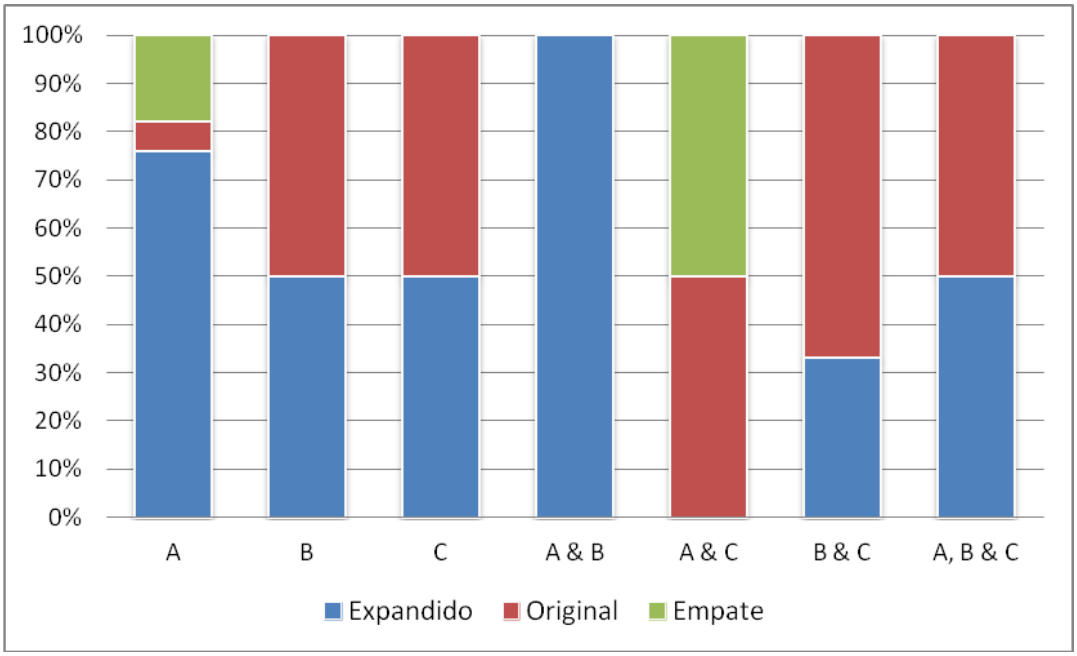


Figura 44 – Precisão Total Agrupada

A segunda métrica, correlação de *ranking* (Figura 45), apresentou melhores resultados para a consulta original em relação à consulta expandida em ambas as

distribuições (as consultas “Original 1” e “Expandido 1” consideraram a distribuição proposta na coluna Correlação’ da Tabela 7, enquanto o valor de “Original 2” e “Expandido 2” consideraram a distribuição proposta na coluna Correlação’’ da mesma tabela). Pode-se observar também que, Correlação’ em relação à Correlação’’, houve leve piora na medida de correlação para as consultas expandidas e melhora de aproximadamente 10% entre as consultas originais.

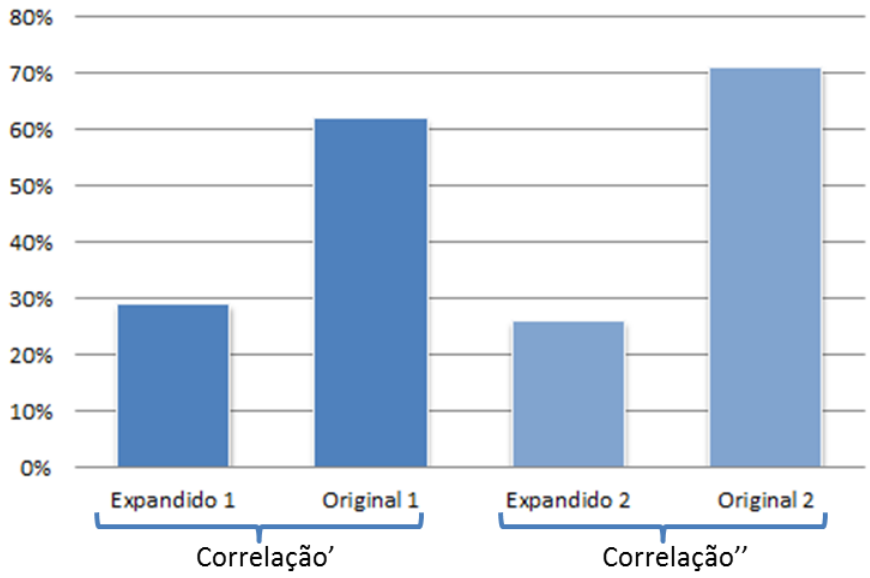


Figura 45 – Correlação de *Ranking*

Na última métrica, comprimento da busca (Figura 46), foi verificada a superioridade das consultas expandidas em relação às consultas originais. Em Expandido 1 (profundidade até se obter o primeiro documento com avaliação de quatro ou cinco estrelas) e Expandido 2 (profundidade até obter-se dois documentos contínuos com avaliação de quatro ou cinco estrelas), 41% dos resultados obtidos com as consultas expandidas foram melhores do que as obtidas com as consultas originais.

Figura 46 – Comprimento da Busca

6.5.3. Análise Qualitativa

A análise qualitativa deste trabalho considera a possível influência do comportamento humano nos resultados obtidos, e, portanto, não há como generalizar os resultados. Para colher indícios de comportamento e uso, perguntou-se a este grupo quais as impressões ao utilizar o protótipo. A maioria dos alunos achou que a ferramenta os ajudou a encontrar resultados relevantes na *Web*. O ponto negativo foi a demora na resposta enviada pelo agente. Segue a íntegra dos principais comentários.

- ALUNO 1: "A ferramenta é muito interessante. O que eu menos gostei foi o tempo necessário até receber os resultados da minha pesquisa, tendo em vista que hoje em dia um fator como este pode atrapalhar a aceitação da ferramenta."
- ALUNO 2: "Poderia existir uma dessas para as aulas de pesquisa, mas já com as estrelas das opiniões de pessoas que já passaram por aquela página de acordo com o que foi pesquisado (pelo grupo). Dessa forma, evitaríamos informações que não são relevantes ou até mesmo as que não tem nada a ver com o assunto pesquisado, o que no caso, aconteceu muito quando eu pesquisava para as aulas."

- ALUNO 3: “A ferramenta é boa, embora eu tenha demorado a receber as respostas das minhas pesquisas. A maioria dos *links* q foram apresentado pra mim me ajudaram a obter informação sobre o que eu queria.”
- ALUNO 4: “A ferramenta funcionou muito bem, apesar de apresentar alguns erros no início e de demorar para responder esses erros foram rapidamente corrigidos, a ferramenta passou a responder quase que imediatamente, os termos dos grupos "A", "B" e "C", apesar de não serem todos totalmente relacionados ao termo pesquisado, a maioria se mostrou útil ao resultado final da pesquisa.”

6.5.4. Considerações Finais

Devem ser consideradas algumas intervenções durante a realização das atividades. A primeira refere-se ao bigrama “Lojas Americanas”, identificado logo no início das discussões (Figura 47). Embora o protótipo tenha acertado ao incluir “lojas americanas” à base contextual enriquecida, esta, de fato, tem pouco a ver com o tema da discussão. Como este casamento aconteceu logo no início das discussões, com o agravante de trazer consigo mais de 230 instâncias[74] irmãs, houve uma explosão de artigos pouco relacionados, de diversas lojas e empresas do Brasil. Portanto, resolveu-se excluir manualmente da base de conhecimento enriquecida todas as instâncias relacionadas a “Lojas Americanas”.

[74] dcterms:subject: http://dbpedia.org/page/Category:Companies_of_Brazil

Figura 47 – Bigrama "Lojas Americanas"

A segunda intervenção foi remover a palavra "TI" da lista de *stopwords*, visto que uma consulta informou somente o termo "TI" e o processamento da consulta as removeu, o que tornou a consulta vazia. Como o código do protótipo não estava preparado para tratar consultas vazias, houve erro de execução não tratado. Após intervenção, executou-se a consulta novamente, entretanto o agente "parou de responder" por cerca de 45 minutos.

No entanto, apesar das intervenções, alguns pontos positivos podem ser observados. O primeiro, apesar de "lojas americanas" não fazer parte do assunto definido para a aula, esta se associou corretamente ao conceito (em dados abertos) de uma "empresa brasileira do segmento de varejo". Também não se constatou nenhuma ambiguidade nos demais bigramas e trigramas deste estudo de caso, como poderia ocorrer com o trigrama "Pão de Açúcar", que poderia representar um ponto turístico ou uma empresa brasileira. Ainda, todos os bigramas e trigramas encontrados no texto, embora não sejam 100% aderentes à temática proposta, relacionam-se de alguma maneira às discussões e pode ser útil à busca, pelo simples fato de terem sido mencionados e encontrados em dados abertos.

Capítulo 7 – Conclusão

Este capítulo apresenta uma visão conclusiva sobre esta dissertação, suas dinâmicas e argumenta a importância de sistema de recuperação de informação em ambientes de aprendizagem colaborativa. Nesses ambientes, a descoberta de documentos relevantes à aprendizagem e relacionados à discussão podem contribuir para o desenvolvimento e disseminação de novos conhecimentos.

7.1. Discussão sobre as propostas

O objetivo desta dissertação foi melhorar a relevância dos resultados das buscas à *Web* a partir do tratamento de informação contextual obtida em mensagens de grupos de redes sociais online. Verificou-se a viabilidade da proposta de recuperação de informação contextual em um grupo de uma plataforma de rede social mundialmente conhecida (*Facebook*). Usou-se um conjunto de bibliotecas *open source* para interagir com *Facebook* e explorar com sucesso os recursos do grupo.

Para melhorar a relevância dos resultados, organizou-se o trabalho em duas propostas que foram desenvolvidas e avaliadas. Estas propostas usaram a técnica de expansão de consultas para a atividade de busca na *Web* e empregaram os mesmos critérios de avaliação. Cada proposta relaciona-se a uma arquitetura, protótipo e estudo de caso distinto. Diferenciam-se quanto à origem dos dados, técnicas usadas para a geração do contexto e forma em que os termos foram extraídos.

O primeiro estudo é uma extensão do trabalho de segmentação e agrupamento proposto originalmente por Prates (2011). Para a modelagem do contexto usou-se documentos de aprendizagem (material fornecido pelo professor) e também mensagens de discussão (*posts*, *links* e comentários que resultam da interação dos participantes). A combinação destas duas fontes de informação gerou três contextos distintos: (i) documentos, (ii) mensagens e (iii) documentos e mensagens. O estudo de caso mostrou que a expansão da consulta é interessante em um ambiente educacional para melhorar os resultados de busca. Houve melhoria na precisão total dos resultados, principalmente aqueles relacionados ao contexto gerado a partir das mensagens de discussão (10% superior ao contexto gerado a partir de notas de aula). Os resultados melhores foram o ponto crucial para a segunda proposta, que focou somente nas discussões. Para adequar-se ao requisito de atendimento de consultas durante as discussões, uma nova proposta foi elaborada e baseou-se nas observações e resultados da primeira proposta (primeiro estudo de caso), descritas a seguir:

- Se tivessem dúvidas durante a realização das atividades, os alunos deveriam usar motores de busca (como o Google) externamente ao ambiente de discussões. Os alunos só tiveram acesso ao protótipo de busca ao final das discussões;
- Durante a etapa de discussões, observou-se troca de mensagens entre os participantes, visto que estes estavam ocupados "caçando" *links* relacionados ao tema proposto em motores de busca na *Web*, para então publicá-los no grupo de discussões. Observou-se então um grande esforço individual e, como resultado deste esforço, muitos *links* repetidos.
- A fase de extração de informação considerou o conteúdo textual de cada *link* para a geração do contexto. Entretanto, se os *links* são enviados ao

grupo porque foram considerados relevantes, então o protótipo possivelmente sugeriria termos para a expansão de consultas que levariam a documentos iguais ou muito semelhantes aos publicados pelos alunos (portanto relevantes).

- A extração dos termos (a partir do *corpus* composto por documentos e discussões) realizou-se somente após o término da dinâmica, ou seja, os termos sugeridos são estáticos e, portanto, são os mesmos para diferentes momentos de consultas dos usuários. Para transpor este problema, uma proposta para automatizar o processo de geração do contexto e permitir que as buscas fossem realizadas durante as discussões esbarrou na complexidade dos algoritmos de segmentação e agrupamento e, portanto, poderia não atender a requisitos de tempo para as buscas (FRITZEN *et* al., 2011);
- Muitas avaliações (50) para definir a relevância dos resultados poderiam comprometer a qualidade das discussões.

Com base nas observações feitas no primeiro estudo de caso, pensou-se em uma nova abordagem onde os alunos pudessem: (i) usar o mesmo ambiente das discussões para a realização das buscas; (ii) priorizar as discussões, ressaltando o caráter colaborativo; (iii) usar somente informação das mensagens para geração do contexto e (iv) simplificar o processo de avaliação.

O segundo estudo investigou as ressalvas geradas a partir da avaliação da primeira arquitetura. Essa nova arquitetura foi concebida para fornecer um mecanismo de busca adaptativo, que tornasse as buscas contextualizadas à medida que novas discussões evoluíssem no grupo. Os termos foram sugeridos de acordo com a similaridade

semântica latente dos termos que compõem a expressão de busca, enquanto na proposta anterior todos os termos sugeridos são idênticos, independente da expressão de busca.

A abordagem utilizada para a construção coletiva do conhecimento considerou somente discussões e troca de ideias com os colegas. Neste novo estudo, observou-se grande interação entre os participantes da dinâmica, que aproveitaram os cinco (5) dias para discutir assuntos relacionados ao tema proposto. No primeiro estudo de caso, apesar de contar apenas com 1 hora de duração, foram criados 53 tópicos de discussão (sendo 83% destes *links*) e 69 comentários, o que implica em uma média de 1,3 comentários por tópico. Já o segundo estudo de caso foram contabilizados 29 tópicos de discussão e 285 mensagens de resposta, uma média de, aproximadamente, 10 respostas por tópico, o que denota a superioridade relacionada à média de mensagens trocadas em torno de um mesmo tópico, ou seja, houve mais discussão sobre cada assunto, para o segundo estudo de caso. Outro ponto positivo do segundo estudo de caso é relacionado à colaboração entre colegas, visto que 66% dos alunos, após a execução de uma busca em uma linha de discussão, voltaram ao grupo e contribuíram com um comentário sobre a informação recuperada, o que mostra a aplicabilidade desta proposta na educação. A correlação de Pearson (KENDALL e STUART, 1973) também foi aplicada a matriz [número de consultas; número de comentários] e obteve valor de correlação 0,79, ou seja, existe uma relação direta entre o número de consultas e o número de comentários por assunto. Por fim, a profundidade média (contagem do número de respostas até que se faça a primeira busca) foi de, aproximadamente, 1,6 e a média de buscas por *threads* de discussão foi de, aproximadamente, duas buscas.

Ressalta-se também a importância da qualidade das mensagens para a correta definição do contexto e valoriza-se o papel do professor em não deixar que as

discussões desvirtuam-se da temática definida no planejamento para a dinâmica e, consequentemente, melhore a aderência dos termos sugeridos as consultas.

Como as mensagens de discussão (desconsiderando-se *links*) apresentaram no primeiro estudo de caso pouco conteúdo textual, propôs-se o enriquecimento semântico das mensagens, que consiste em adicionar *corpus* textual de artigos relacionados às discussões, com o uso de uma ontologia de enciclopédia, a fim de melhorar o contexto do domínio das discussões e, consequentemente, melhorar também a qualidade dos termos sugeridos para expansão das consultas. Para as consultas, propôs-se um agente de interface que assiste os alunos a pesquisar documentos na *Web*. Esse assistente fornece ajuda aos alunos, ao identificar os principais termos do contexto da conversa e relacionados à sua expressão de busca, dentro de um conjunto muito maior de termos. O sistema sugere palavras-chave relevantes a partir das discussões, que podem ser combinadas pelo usuário. As consultas, quando expandidas com termos do contexto enriquecido em dados abertos, apresentaram resultados significativamente melhores que os obtidos com a expansão do contexto sem enriquecimento ou pela consulta original. Com isso, este trabalho sugere que a modelagem do domínio a partir de discussões pode situar o aluno nesse domínio. A segunda proposta de arquitetura obteve melhoria de 62% na métrica precisão total para os resultados de consulta expandidos, que, comparado à métrica de precisão total / aba “expansão de contexto” da primeira arquitetura, foi 17% melhor para o contexto obtido a partir de mensagens de discussão, e 37% melhor que o contexto obtido a partir de notas de aula.

Por fim, esta dissertação defendeu a criação do contexto de domínio de maneira automática, ou seja, com o uso de recursos e tecnologias existentes, que independam de esforço humano adicional para a modelagem deste domínio. A modelagem de novas ontologias foi descartada. A engenharia de ontologias tem custo elevado e exige

dedicação de especialistas e tempo para a modelagem. Criar modelos de representação e realizar marcações (anotações semânticas) a partir desses modelos é uma tarefa trabalhosa que pode requerer dedicação de especialistas, embora possa apresentar bons resultados se bem realizada.

7.2. Contribuições

Este trabalho apresentou duas propostas de recuperação de informação contextual na *Web* em grupos de discussão: Captura do Contexto a Partir de Discussões e Enriquecimento de Termos das Discussões. Ambas as propostas exigiram pouco esforço do especialista para a definição do contexto. Para cada proposta desenvolveu-se um protótipo de sistema de informação, usado para a realização de seus respectivos estudos de caso. Os resultados aferidos, relacionados à primeira proposta (Captura do Contexto a Partir de Discussões) apresentaram certas inconsistências para a geração de contextos dinâmicos e foram úteis para a criação de uma nova proposta (Enriquecimento de Termos das Discussões). Esta nova proposta representa o principal campo de contribuição desta pesquisa e, dentre as principais contribuições, podem ser citadas:

- Enquanto outras abordagens para modelagem do contexto partem de *corporas* textuais previamente definidas, a presente proposta independe de qualquer tipo de conteúdo pré-existente para representar o contexto, que é desenvolvido e moldado exclusivamente a partir das discussões. Também se propôs o enriquecimento do contexto das mensagens em dados abertos, a fim de ampliar a completude terminológica das mensagens, visto que estas, normalmente, são compostas por poucos termos. Este contexto, chamado de contexto enriquecido, sugere termos relacionados à expressão de busca do usuário para a expansão de consultas em grupos de redes sociais, para apoiar as buscas na *Web*.

- Melhores resultados das consultas expandidas, obtidos pela métrica de precisão total em relação à consulta original, foram apresentados nas consultas que usaram termos extraídos do contexto enriquecido (“Grupo A”, “Grupo B” e “Grupo A e B”). Apesar de a amostragem ser pequena para finalidade de resultados conclusivos, a expansão com termos do “Grupo A” (rótulos da *DBpedia*) apresentou resultados promissores frente aos demais.
- Desenvolvimento de uma arquitetura baseada em agentes para a contínua atividade de extração de mensagens e geração do contexto de domínio.

7.3. Limitações da abordagem e trabalhos futuros

A falta de mecanismos de controle para desambiguação dos bigramas e trigramas que são incluídos a base contextual poderia ser melhor explorada, como o problema das “lojas americanas”, um bigrama fora do contexto da aula, que, expandido, foi maior que o *corpus* pré-existente. Algoritmos de agrupamento poderiam detectar artigos “*outlier*” e descartá-los de forma automática ou semiautomática do contexto enriquecido, mediante confirmação dos alunos ou do professor. Outra solução poderia simplesmente adotar uma regra para manter a qualidade do *corpus* atual, por exemplo, a verificação de novas instâncias enriquecidas não poderiam excederem um limite pré-estabelecido para a quantidade atual de instâncias presentes no *corpus* enriquecido.

Outra limitação é a ausência de garantias para a geração do contexto enriquecido, a partir do casamento entre os bigramas e trigramas das mensagens e instâncias em dados abertos (*DBpedia*), pois o domínio discutido pode não existir neste repositório.

Dentre os possíveis trabalhos futuros, destacam-se:

- Testar outras formas de organização e visualização dos termos. Isso visa facilitar o entendimento e consequentemente a seleção dos termos pelos usuários. Como

exemplo, os termos poderiam ser agrupados de forma hierárquica, conforme sugerido em (JOHO *et* al., 2004);

- Personalizar a geração do contexto por aluno, selecionando apenas as *threads* de mensagens que este tenha alguma participação (interesse). Pode ser visto como um subconjunto do contexto de domínio atribuído a cada aluno.
- Considerar pesos diferenciados para a geração dos termos, de acordo com (i) a *thread* em que a busca foi realizada e (ii) priorização de laços sociais entre o solicitante da consulta e sua rede.
- Avaliar a interação de cada aluno com o sistema e com os demais alunos e investigar se houve melhoria na aprendizagem (questão em aberto).
- Avaliar os termos sugeridos a cada consulta, possibilitando pesos diferenciados aos termos de acordo com as avaliações recebidas em consultas passadas.
- Usar as relações da ontologia da *DBpedia*[75], se existirem conceitos relacionados ao domínio em questão. O presente trabalho considerou somente instâncias, visto que, a maior parte dos artigos de "Ciência da Computação" relaciona-se a um conceito geral (*thing*). Cabe destacar que a ontologia da *DBpedia* está em constante evolução pela comunidade e outros assuntos já possuem alguma modelagem conceitual, como "Filmes", "Automóveis", "Política" e "Geografia".
- Usar outros modelos para representação do conhecimento, tais como ontologias ou tesauros, como fonte de informação para a garantia de qualidade do contexto de domínio. Para alinhar-se aos objetivos desta pesquisa de esforço mínimo desprendido por especialistas, devem-se buscar modelos existentes, que contemplem certo grau de consenso de uma comunidade, tal como a ontologia

[75] http://dbpedia.org/ontology/

para a ciência da computação [76] . Esta ontologia contém conceitos e relacionamentos que abrangem as mais variadas disciplinas pertencentes ao domínio da ciência da computação e já foi explorada por Paula (2010).

- Utilizar outros algoritmos para relacionar os termos da consulta com os documentos que compõem o contexto enriquecido, como a projeção aleatória (SAHLGREN, 2005) e a indexação aleatória reflexiva (COHEN *et* al., 2010).
- Realizar um novo estudo de caso com uma amostragem maior e permitir a solicitação de consultas logo após o primeiro enriquecimento (neste estudo, esperou-se um dia para a geração do contexto enriquecido).
- Realizar novos estudos em outros domínios que não informática, tais como História, Música, Religião etc.
- Realizar outros estudos com outros perfis de usuários, como grupos de pesquisadores e comunidade de prática em organizações (WENGER, 1998).
- Habilitar o suporte a outros ambientes, como o *Google+* ou *Moodle* por exemplo.

[76] http://www.distributedexpertise.org/computingontology

Referências

ALEXA., The top 500 sites on the web, 2012. http://www.alexa.com/topsites.

ALLAN, J.,CARTERETTE, B., LEWIS, J., "When will Information Retrieval be "Good Enough?"". *Proceedings of the 28th annual ACM SIGIR*, 2005.

ALUISIO, S., PELIZZONI, J., MARCHI, A., OLIVEIRA, L., MANENTI, R., MARQUIAFAVEL, V., "An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese", In *Propor*, Pages 110–117, 2003.

AMBRÓSIO, A.P., SILVA, L.O., NETO, V.G., "Automatic Retrieval of Complementary Learning Material for Slide Presentations". *International Conference on Interactive Computer Aided Blended Learning (ICBL 2009)*. Retrieved from: http://www.valdemarneto.com/pdfs/artigoICBL.pdf.

ANTONIOU, G., HARMELEN, F., "A Semantic Web Primer", 2nd ed., 84 Massachusetts, MIT Press, 2004.

AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., IVES, Z., "Dbpedia: A nucleus for a *Web* of open data", *The Semantic Web*, *4825*(Springer), 722-735. Springer, 2007.

BAKER, R., YACEF, K., "The State of Educational Data Mining in 2009: A Review and Future Visions*". Journal of Educational Data Mining (JEDM)*, 1(1), 3–17, 2009.

BARRETO, F., BRANCO, A., FERREIRA, E., MENDES, A., NASCIMENTO, M., NUNES, F., SILVA, J., "Linguistic Resources and Software for Shallow Processing", in Oliveira, Fátima e Joaquim Barbosa (orgs.) *Actas do XXI Encontro Nacional de Linguística*, Lisboa, Associação Portuguesa de Linguística, pp. 203-217, 2006.

BAZIRE, M., BRÉZILLON, P., "Understanding Context Before Using It". *In Modeling and Using Context, 5th International and Interdisciplinary Conference (CONTEXT 2005)*, volume 3554 of Lecture Notes in Computer Science, pages 29–40. Springer, July 2005.

BCG - Clicks Grow Like BRICS: G-20, Internet Economy to Expand at 10 Percent a Year Through 2016. 2012. Disponível em: http://www.bcg.com/media/PressReleaseDetails.aspx?id=tcm:12-100468.

BERNERS-LEE, T., HENDLER, J., LASSILA, O., The Semantic *Web*. (A. Gómez-Pérez, Y. Yu, & Y. Ding, Eds.) *Scientific American*, *284*(5), 34-43. Citeseer, 2001. Retrieved from http://www.nature.com/doifinder/10.1038/scientificamerican0501-34.

BERRY, M. W., DRMAC, Z., JESSUP, E.R. "Matrices, vector spaces, and information retrieval". *Siam Review*, 41(2): página 335-362, 1999.

BEKKERMAN, R., ALLAN, J., "Using Bigrams in Text Categorization". *Center for Intelligent Information Retrieval (CIIR)*, Technical Report IR-408, 2004.

BELLIFEMINE, F. L.; CAIRE, G.; GREENWOOD, D. "Developing Multi-Agent Systems with Jade". [S.l.]: Wiley, 2007. (*Wiley Series in Agent Technology*)

BHOGAL, J., MACFARLANE, A., SMITH, P., "A review of ontology based query expansion". *Information Processing & Management*, 43(4), 866-886. Elsevier, 2007. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0306457306001476

BICK, ECKHARD. "Structural Lexical Heuristics in the Automatic Analysis of Portuguese". *11th Nordic Conference on Computational Linguistics*, Copenhagen, 1998. p.44-56.

BING, "Bing Help Home", abr. 2012. Disponível em: http://onlinehelp.microsoft.com/en-us/bing/.

BITTENCOURT, I. I., Modelos e Ferramentas para a Construção de Sistemas Educacionais Adaptativos e Semânticos, 2009. (Tese de Doutorado).

BITTENCOURT, I. I; COSTA, E., "Modelos e Ferramentas para a Construção de Sistemas Educacionais Adaptativos e Semânticos". Revista Brasileira de Informática na Educação. 19(01): 85 - 98. 2011.

BITTENCOURT, I. I., "Plataforma para Construção de Ambientes Interativos de Aprendizagem baseados em Agentes". Universidade Federal de Alagoas, 2006. (Mestrado)

BIZER, C., CYGANIAK, R., HEATH, T. "How to Publish Linked Data on the Web", 2007. Disponível em http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

BIZER, C., HEATH, T., BERNERS-LEE, T., "Linked Data - The Story So Far". (T Heath, M. Hepp, & C Bizer, Eds.) *International Journal on Semantic Web and Information Systems*, *5*(3), 1-22. Elsevier, 2009. Retrieved from http://www.citeulike.org/user/omunoz/article/5008761.

BRADSHAW, J. M., "Introduction to software Agents", *Software Agents*, ed. Bradshaw, J. M. Menlo Park, Calif.: AAAI Press, 490 p. ISBN 0-262-52234-9, 1997.

BRAZ, L. M., SERRÃO, T., PINTO, S. C. C. S., CLUNIE, G., "Um Mecanismo para a Integração entre o LMS Moodle e o Site de Redes Sociais Facebook", *Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE)*, Aracaju, 2011.

BREITMAN, K., “Web Semântica: a Internet do futuro”. Rio de Janeiro: LTC, 2005.

CAIRE, G., "JADE PROGRAMMING FOR BEGINNERS". *TILAB*. 2009. Disponível em http://jade.tilab.com/doc/tutorials/JADEProgramming-Tutorial-for-beginners.pdf.

CARPINETO, C., ROMANO, G., “A Survey of Automatic Query Expansion in Information Retrieval”. *ACM Computing Surveys*, 44(1), 1:1-1:50, 2012.

CHANANA, V., GINIGE, A., MURUGESAN, S., “Improving information retrieval effectiveness by assigning context to documents”. *International Symposium on Information and Communication Technologies (ISICT 2004)* (pp. 2239-2244). IEEE Press, New York, 2004.

CHEN, L., SYCARA, K., “WebMate: A Personal Agent for Browsing and Searching *”. *Knowledge Acquisition*, 132-139, 1997.

CHIGNELL, M. H., GWIZDKA, J., BODNER, R. C. “Discriminating meta-search: A framework for evaluation”, *Information Processing and Management: an International Journal*, v. 35, issue 3, pp-337-362, 1999.

COHEN, T., SCHVANEVELDT, R., & WIDDOWS, D., “Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections”, *Journal of Biomedical Informatics*, *43*(2), 240-256. Elsevier Inc. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19761870.

COMSCORE.,” It’s a Social World: Top 10 Need-to-Knows About Social Networking and Where It’s Headed”, dez. 2011. Disponível em: http://www.comscore.com/por/Press_Events/Presentations_Whitepapers/2011/it_is_a_social_world_top_10_need-to-knows_about_social_networking.

CONRADO, M. S., MARCACINI, R. M., MOURA, M. F., REZENDE, S. O. “O Efeito do uso de Diferentes Formas de Geração de Termos na Compreensibilidade e Representatividade dos Termos em Coleções Textuais na Língua Portuguesa”. In: *The 7th Brazilian Symposium in Information and Human Language Technology (STIL)* - II International Workshop on Web and Text Intelligence (WTI): ICMC - USP, 2009. p.1 – 10.

COOPER, W. S. “Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems”, *Journal of American Society of Information Science*, v. 19, issue 1, pp-30-41, 1968

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., HARSHMAN, R., “Indexing by latent semantic analysis”, *Journal of the American Society for Information Science*, *41*(6), 391-407. Citeseer, 1990. Retrieved from http://doi.wiley.com/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9

DELOACH , S. A., “Analysis and Design using MaSE and agentTool”, 2001.

DEY, A. K., ABOWD, G. D. “Towards a better understanding of context and contextawareness”. *In: Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, pp. 304-307. Springer-Verlag, 1999.

DEY, A. K., “Understanding and Using Context”. *In Personal Ubiquitous Computing*, 5(1):4–7, 2001.

DOTTA, S., “Uso de uma Mídia Social como Ambiente Virtual de Aprendizagem”. *Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE) - XVII WIE*, Aracaju, 2011.

ESTOPÀ BAGOT, R. “Extracció de terminologia: elements per a la construcció d’un SEACUSE (*Sistema d’Extracció Automàtica de Candidats a Unitats de Significació Especialitzada*)”. Tese de Doutorado. Universidade Pompeu Fabra, 1999.

EXPERIAN HITWISE., “Buscas com uma palavra são maioria”, dez. 2011. Disponível em: http://www.serasaexperian.com.br/release/noticias/2011/noticia_00728.htm.

FERNANDES, E. L. R., MILIDIÚ, R. L., SANTOS, C. N., “Portuguese language processing service”. *Computer, 2009*. Retrieved from http://wwwconference.org/www2009/pdf/submissions/wwwiberoamerica09_submission_1.pdf

FERNANDEZ-LOPEZ, M., CORCHO, O. “Ontological Engineering*”. Advanced Information and Knowledge Processing Series.* 420 p. 1st ed. Springer, 2004.

FERNEDA, E, “Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação”. *Tese de doutorado defendida na Escola de Comunicação e Artes da Universidade de São Paulo.* USP. 2003.

FININ, T., LABROU, Y., KQML as an agent communication language. . J.M. In: Bradshaw (ed.), Software Agents, pp. 291-316. Cambridge, MA, 1997.

FIPA specification SC00061G: FIPA ACL Message Structure Specification, 2002, http://www.fipa.org/specs/fipa00061/SC00061G.html.

FOLTZ, P., “Latent semantic analysis for text-based research. Behavior Research Methods”, *Instruments and Computers*, 28(2):197-202, 1996.

FRANKLIN, S., GRAESSER, A., “Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents”. (J. Müller, M. Wooldridge, & N. Jennings, Eds.) *Intelligent Agents III Agent Theories Architectures and Languages*, *1193*(2), 21-35. Springer-Verlag, 1996. Retrieved from http://www.springerlink.com/index/w5m51l674402vr07.pdf.

FRANKLIN, T., HARMELEN, M., “*Web* 2.0 for Content for Learning and Teaching in Higher Education”., *Teaching in Higher Education*, *2008*(May), 1-29. JISC, 2007. Retrieved from http://ie-repository.jisc.ac.uk/148/.

FRITZEN, E., SIQUEIRA. S. W. M. , ANDRADE, L. C. V., “An agent-oriented system for contextualized web queries”., *Proceedings of the IADIS International Conference on WWW/Internet 2011*. P. 479-483. Rio de Janeiro, Brasil. 2011.

GATTI, L. A. C.,, "Uma Arquitetura Baseada em Contexto de Atividades para Gestão de Conhecimento em Processos de Trabalho”, 2009. Dissertação de M.Sc., PPGI/UNIRIO, Rio de Janeiro, RJ, Brasil.

GATTI, L. A. C., SANTORO, F. M., NUNES, V. T., “An agent-based architecture for knowledge management in context-aware business processes”. *Computer Supported Cooperative Work in Design (CSCWD), 2010*: 318-323.

GLUZ, J. C., VICCARI, R. M., “Uma Ontologia OWL para Metadados IEEELOM, DublinCore e OBAA”. *Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE) - XVII WIE*, Aracaju, 2011, p. 1-10.

GOOGLE, “Ajuda da Pesquisa na Web”, abr. 2012. Disponível em: http://www.google.com.br/support/websearch/?hl=br

GRIFFITHS, J. R., BROPHY, P., “Student Searching Behavior and the Web: Use of Academic Resources and Google”, *Library Trends*, v. 53 n. 4, pp.539-54, 2005.

GRUBER, T.,, “A translation approach to portable ontologies”, *Knowledge Acquisition*, 5(2):199-220, 1993.

GRUBER, T., “Collective knowledge systems: Where the Social Web meets the Semantic Web”. *Journal of Web Semantics* 6(1), 4–13, 2008.

HEARST, M.A. “TextTiling: Segmenting text into multi-paragraph subtopic passages”, *Computational Linguistics*, pp 33-64, 1997.

HENDERSON-SELLERS, B.; GIORGINI, P., “Agent-Oriented Methodologies”. Hershey; London; Melbourne; Singapore: Idea Group, 2005. 413 p.

HERLI, J. M., (2011). "Composição de Objetos de Aprendizagem com Base em Semiótica”. Dissertação de M.Sc., PPGI/UNIRIO, Rio de Janeiro, RJ, Brasil.

IBOPE,, “Brasileiros caem na rede social”, jan. 2012. Instituto Brasileiro de Opinião Pública e Estatística, 2011. Disponível em: http://www.ibope.com.br/calandraWeb/servlet/CalandraRedirect?temp=5&proj=PortalIBOPE&pub=T&db=caldb&comp=IBOPE+Media&docid=39D1E142AFCFDAF88325782400545EE9

ISOTANI, S., BITTENCOURT, I. I., COSTA, E., MIZOGUCHI, R., Estado da arte em web 3.0: Potencialidades e tendências da nova geração de ambientes de ensino na web. Revista Brasileira de Informática na Educação, 17(1):30–42, 2009.

JANSEN, B. J., SPINK, A., BATEMAN, J., SARACEVIC, T. “Real life information retrieval: A study of user queries on the Web”, *ACM SIGIR Forum*, 32(1), pp 5-17, 1998.

JOHNSON, D., MALHOTRA, V., VAMPLEW, P., “More Effective *Web* Search Using Bigrams and Trigrams”, *Webology*, *3*(4), 1-12, 2006. Retrieved from www.webology.ir/2006/v3n4/a35.html.

JOHNSON R, FOOTE B., “Designing Reusable classes. Journal of Object-Oriented Programming”, 1(2):22?35, June/July 1988.

JOHO, H., SANDERSON, M., BEAULIEU, M. "A study of user interaction with a concept-based interactive query expansion support tool". *In: Proceedings of the European Conference on IR Research*, Springer , pp. 42-56, Apr. 2004.

JURAFSKY, D., & MARTIN, J. H., *Speech and Language Processing*. (A. Kehler, K. Vander Linden, & N. Ward, Eds.) *Intensive Care Medicine* (Vol. 36 Suppl 1, pp. S4-10). Prentice Hall, 2000. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20413954.

KANAAN, G., AL-SHALABI, R., GHWANMEH, S., AND BANI-ISMAIL, B., “Interactive and automatic query expansion: A comparative study with an application on Arabic”. Amer. J. Appl. Sciences 5, 11, 1433–1436, 2008.

KANG, J. W., KANG, H., KO, M., JEON, H. S., & NAM, J., “A Term Cluster Query Expansion Model Based on Classification Information in Natural Language Information Retrieval”. *In: Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence* (pp. 172 – 176), 2010.

KAPLAN, A. M., HAENLEIN, M., Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 53(1), 59-68. doi:10.1016/j.bushor.2009.09.003, 2010.

KEßLER, C., “Context-aware Semantics-based Information Retrieval. 2010.

KELLY, D., GYLLSTROM, K., BAILEY, E. W., “A Comparison of Query and Term Suggestion Features for Interactive Searching”. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* pp. 371-378, 2009.

KENDALL, M.G., STUART, A., “The Advanced Theory of Statistics”, Volume 2: Inference and Relationship. Griffin, 1973.

KERLINGER, F. N. “Metodologia da pesquisa em ciências sociais”. São Paulo; *EDUSP*, 1980.

KITAMURA, Y., MIZOGUCHI, R., "Ontology-based systematization of functional knowledge", *Journal of Engineering Design*, Taylor & Francis, Vol. 15, Number 4, pp. 327-351, August 2004.

KOBAYASHI, M. E. I., TAKEDA, K., "Information Retrieval on the Web", *ACM Computing Surveys*, Vol. 32, No. 2, 144-173, 2000.

KOWALSKI, G., "Information Retrieval Systems: Theory and Implementation". *Kluwer Academic Publishers*, 1997. 282 p.

KRIEGER, M.G.; FINATTO, M.J.B., "Introdução à terminologia. Teoria e Prática". *Editora Contexto*, São Paulo, SP. 223 pp, 2001.

KUIPER, E., VOLMAN, M., TERWEL, J., "The *Web* as an Information Resource in K-12 Education: Strategies for Supporting Students in Searching and Processing Information". *Review of Educational Research, 75*(3), 285-328. Retrieved from http://rer.sagepub.com/cgi/doi/10.3102/00346543075003285

LAKATOS, E. M., MARCONI, M. A.., "Metodologia do trabalho científico: procedimentos básicos, pesquisa bibliográfica, projeto e relatório, publicações e trabalhos científicos" – 6. ed. – São Paulo : Atlas, 2001.

LÉVY, P., "Cibercultura". São Paulo: 34, 1999.

LIKERT, RENSIS, "A Techinique for the Measurement of Attitudes", *Archives of Psychology*, 140: PP. 1-55, 1932.

LIZOTTE, M; MOULIN, B. "A Temporal Planner for Modelling Autonomous Agents". *In: Decentralized A.I.*, 1990.

MAES, P., "Agents that Reduce Work and Information Overload". *Communications of the ACM*, ACM Press, v. 37, n. 7, p. 31-40, jul. 1994.

MAES, P., "Artificial Life Meets Entertainment: Life like Autonomous Agents"., *Communications of the ACM*, ACM Press, v. 38, n. 11, p. 108-114, 1995.

MANNING, C. D., SCHÜTZE, H., "Foundations of Statistical Natural Language Processing*"*, (M. I. T. P. Cambridge, Ed.) *Computational Linguistics* (Vol. 26, pp. 277-279). MIT Press, 1999. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/coli.2000.26.2.277.

MANNING, C.D., RAGHAVAN, P., SCHÜTZE, H., "Introduction to Information Retrieval", *Cambridge University Press*, Cambridge, 2008.

MANSUR, A. F. U., CARVALHO, R. A. D., AND BIAZUS, M. C. V., "Rede de Saberes Coletivos (RESA): Um Ambiente Complexo para Aprendizagem Acadêmica por Meio de Redes Sociais", *Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE) - XVII WIE*, Aracaju, 2011. 1284-1293, 2011.

MATTOS, D. O. P., ROSA+: Uma Extensão do Modelo ROSA com Suporte a Regras e Inferência, 2006. Dissertação (mestrado) – Instituto Militar de Engenharia – Rio de Janeiro.

MCCARTHY, D., KOELING, R.,WEEDS, J., CARROLL, J., "Unsupervised acquisition of predominant word senses". *Computat. Ling*. 33, 4, 553–590, 2007.

MEDELYAN, O., MILNE, D., LEGG, C., WITTEN, I. H., "Mining meaning from Wikipedia". *International Journal of Human-Computer Studies*, 67(9), 716-754, 2009. Elsevier. doi:10.1016/j.ijhcs.2009.05.004.

MIKROYANNIDIS, A., "Toward a Social Semantic *Web", Computer, 2007*. Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4385271.

MIRIZZI, R., RAGONE, A., NOIA, T. D., SCIASCIO, E. D., "Ranking the Linked Data: The Case of DBpedia", 337-354, 2010.

MOENS, M. F., "Information Extraction: Algorithms and Prospects in a Retrieval Context"., *The Information Retrieval Series.* Vol. 21. 1 ed, Softcover, 2006

MORA-SOTO, A., "Collaborative Learning Experiences Using Social Networks. International Conference on Education and New Learning Technologies" *(EDULEARN09), 2009.* Retrieved from: http://uc3m.academia.edu/ArturoMoraSoto/Papers/114887/Collaborative_ Learning_Experiences_Using_Social_Networks

MOTTA, E. N., "Preenchimento Semi-automático de Ontologias de Domínio a Partir de Textos em Língua Portuguesa". Tese de M.Sc., PPGI/UNIRIO, Rio de Janeiro, RJ, Brasil, 2009.

MOTTA, E. N., FERNANDES, E. R., MILIDIÚ, R. L., "F-EXT-WS-2.0: A *Web* Service for Natural Language Processing", *PROPOR 2010*.

MOURA, M. F., NOGUEIRA, B. M., CONRADO, M. D. S., SANTOS, F. F., REZENDE, S. O., "Making good choices of non-redundant n-gram words". *2008 11th International Conference on Computer and Information Technology*, 64-71. Ieee, 2008.

NARAYAN, N., "Advanced Intranet Search Engine", 2009.

NASCIMENTO, N., PIMENTEL, E., AND DOTTA, S., "Humanização do Ensino Mediado por Computador para Possibilitar uma Aprendizagem mais Colaborativa e Intuitiva", 2138-2147, 2011.

NAVIGLI, R., "Word sense disambiguation". (E. Agirre & P. Edmonds, Eds.)*ACM Computing Surveys*, *41*(2), 1-69. ACM, 2009. Retrieved from http://portal.acm.org/citation.cfm?doid=1459352.1459355

NEDJA, N., MOURELLE, L. M., KACPRZYK J., FRANÇA, F. M. G., SOUZA, A. F., "Intelligent Text Categorization and Clustering", Springer ,2008.

NEGROPONTE, N., “The Architecture Machine: Toward a More Human Environment. Cambridge”, *Mass*, MIT Press, 1970.

NETCRAFT, “Web Server Survey”, out. 2011. Disponível em: http://news.netcraft.com/archives/2011/10/06/october-2011-web-server-survey.html.

NORMAN, D. A., “How Might People Interact with Agents? In Software Agents”, ed J. M. Bradshaw. Menlo Park, Calif.: *AAAI Press, 1997.*

OLIVEIRA, E. W., "Segmentação de objetos de aprendizagem e abordagens para sua utilização”. Dissertação de M.Sc., PPGI/UNIRIO, Rio de Janeiro, RJ, Brasil, 2009.

OREILLY, T., “What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software!., *Communications & Strategies*, No. 1, p. 17, First Quarter 2007. Available at SSRN: http://ssrn.com/abstract=1008839.

ORENGO, V., HUYCK, C., “A stemming algorithm for the Portuguese language”., *In: Proceedings of the 8th International Symposium on String Processing and Information Retrieva l* (SPIRE) 2001. pp 186-193.

OSGOOD, C. E., SUCI, G., TANNENBAUM, P., “The measurement of meaning”. *Urbana,* IL: University of Illinois Press, 1957.

PAULA, Á. A. B., “Uma Proposta para Expansão Semântica de Consultas Baseada em Ontologia de Domínio Específico”, 2010. (Dissertação de Mestrado).

PEAT, H. J.; WILLETT, P. “The limitations of term co-occurrence data for query expansion in document retrieval systems”. *Journal of the American Society for Information Science*, v. 42, n. 5, p. 378-383. doi: 10.1002/(SICI)1097-4571(199106)42:5<378::AID-ASI8>3.0.CO;2-8, 1991.

PECHI, D., “Como usar as redes sociais a favor da aprendizagem”., *Revista Nova Escola, 2011*. Retrieved from: http://revistaescola.abril.com.br/gestao-escolar/redes-sociais-ajudam-interacao-professores-alunos-645267.shtml

PINHEIRO, W. A., “An Ontology Based-Approach for Semantic Search in Portals” *In: Anais do 15º International Workshop on Database and Expert Systems Applications (DEXA),* pp.127-131, Zaragoza, 2004.

BRÉZILLON J., POMEROL, P., “Modeling and using context for system development: Lessons learned from experience”, 1-31, 2001. Retrieved from http://www-sysdef.lip6.fr/~brezil/Pages2/Publications/BP2001.pdf

PORTER, J., “Designing for the Social Web”. *New Riders, 2008*.

PRATES, J. C., “Técnicas de segmentação e agrupamento aplicadas a recursos predeterminados para contextualizar buscas na Internet”. Dissertação de M.Sc., *PPGI/UNIRIO*, Rio de Janeiro, RJ, Brasil, 2011.

PRATES, J. C., SIQUEIRA, S. W. M., "Contextual Query based on Segmentation and Clustering of Selected Documents for Acquiring Web Documents for Supporting Knowledge Management". *Americas Conference on Information Systems (AMCIS)*, AIS Electronic Library (pp. 1-9), 2011a.

PRATES, J. C., SIQUEIRA, S. W. M., "Using educational resources to improve the efficiency of Web searches for additional learning material", *In: IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 563-567), 2011b.

PRATES, J. C., FRITZEN, E., SIQUEIRA, S. W. M., ANDRADE, L. C. V., BRAZ, M. H. L. B., "Improving the Efficiency of Web Searches in Collaborative Learning Platforms". *In: World Summit on the Knowledge Society (WSKS 2011)* (pp. 1-9), 2011.

PRATES, J. C., FRITZEN, E., SIQUEIRA, S. W. M., ANDRADE, L. C. V., BRAZ, M. H. L. B., "Contextual Web Searches in Facebook using Learning Materials and Discussion Messages". *In: Computers in Human Behavior (CHB*), 2012.

PREECE, J., ROGERS, Y., SHARP, H., "Design de interação: além da interação homem-computador"., *Artmed Porto Alegre RS*. Bookman, 2005.

PRUD'HOMMEAUX, E., SEABORNE, A. "SPARQL query language for RDF". *W3C*, 2008.

RENSING, C., SCHOLL, P., BÖHNSTEDT, D., STEINMETZ, R., "Recommending and finding multimedia resources in knowledge acquisition based on Web resources". *Proceedings of 19th International Conference on Computer Communications and Networks (pp. 1-6) IEEE eXpress Conference Publishing, 2008.* Retrieved from ftp://ftp.kom.tu-darmstadt.de/papers/RSBS10.pdf.

RICHARDSON, W., "Blogs, Wikis, Podcasts and Other Powerful *Web* Tools for the Classroom". (2nd ed). *Corwin Press*, Thousand Oaks, 2009.

RIJSBERGEN, C. J. V., CRESTANI F., LALMAS M., "Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information". *Kluwer Academic Publishers*, 1998.

RODRIGUES, R., ASNANI, K., "Concept Based Search Using LSI and Automatic Keyphrase Extraction", In Proceedings of the 2010 3rd International Conference on Emerging Trends in Engineering and Technology (ICETET '10). IEEE Computer Society, Washington, DC, USA, 573-577. DOI=10.1109/ICETET.2010.100 http://dx.doi.org/10.1109/ICETET.2010.100, 2010.

ROMERO, C., VENTURA, S., PECHENIZKIY, M., BAKER, R.S.J.d. (Eds.), "Handbook of Educational Data Mining". CRC Press, 2012.

ROMERO, C. VENTURA, S., "Preface to the Special Issue on Data Mining for Personalised Educational Systems". *User Modeling and User-Adapted Interaction.* 21(1-2), 1-3, 2011.

RUSSELL, S., NORVIG, P., “Artificial Intelligence: A Modern Approach” Prentice-Hall, 1995.

RUTHVEN, I., “Re-examining the potential effectiveness of interactive query expansion”. *In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM Press, 213–220, 2003.

SAHLGREN, M., “An introduction to random indexing”. *In Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, 2005.*

SALTON, G., “Automatic text processing: The transformation, analysis, and retrieval of information by computer”. Addison-Wesley, 1989.

SANTOS, V., “Uma Arquitetura Suportada por Busca Semântica para Recuperação de Fontes de Informação em Repositórios de Metadados”, 2011. Tese de M.Sc., PPGI/UNIRIO, Rio de Janeiro, RJ, Brasil.

SARAWAGI, S., “Information Extraction”. Now Publishers Inc, 2008.

SCHEUER, O., MCLAREN B. M., “Educational Data Mining”. In: *N. M. Seel (Ed.), Encyclopedia of the Sciences of Learning* (pp. 1075-1079). Springer, New York, 2012.

SIQUEIRA, S. W. M., “EDUCO : Modelando Conteúdo Educacional EDUCO : Modelando Conteúdo Educacional”, 2005. PUC-Rio. Doutorado.

SMEATON, A. F., “Information Retrieval : Still Butting Heads with Natural Language Processing” (M. T. Pazienza, Ed.) *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, (1299), 115-138. Springer, 1997. Retrieved from citeseer.ist.psu.edu/smeaton97information.html

SMRZ, P., SCHMIDT, M., “Information extraction in semantic wikis. Proceedings of the 4th Workshop on Semantic Wikis, European Semantic Web Conference”. Citeseer, 2009. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.6810&rep=rep1&type=pdf

SPINK, A., JANSEN, B.J. “A study of Web search trends”, Webology, 1(2), article 4, 2004. Available at: http://www.webology.ir/2004/v1n2/a4.html.

SPINK, A., WOLFRAM, D., JANSEN, M. B. J., AND SARACEVIC, T., "Searching the web: The public and their queries", *Journal of the American Society for Information Science and Technology, 2001.*

STEELE, R., “Techniques for Specialized Search Engines”, *In Proceedings of Internet Computing*. Las Vegas. 2001.

STEEN, M. V., TANENBAUM. A. S.. “Distributed Systems - Principles and Paradigms”. Prentice Hall, 2002.

SU, L. T., CHEN, H. L., DONG, X. Y., "Evaluation of Web-based search engines from an end-user's perspective: A pilot study". *In: Proceedings of the 61st Annual Meeting of the American Society for Information Science*, Pittsburgh, PA., pp-348-361, 1998.

TANG, M.C., SUN, Y. "Evaluation of Web-Based Search Engines Using User-Effort Measures", *LIBRES Research Electronic Journal*, 13(2), 2003.

THOMPSON, J., "Is Education 1.0 Ready for Web 2.0 Students?", 2006, Education.

VARELLA, A. N., "COOPRACTICE – Comunidades de Prática Virtuais Apoiadas por Ontologias", Dissertação de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2007.

VENTURA, J., "Extracção de Unigramas Relevantes", 2008. Retrieved from http://run.unl.pt/handle/10362/1786, (Mestrado).

VIEIRA, R., "Textual Co-reference annotation: a Study on Definite Descriptions". *VIII Congresso da Sociedade Argentina de Lingüística*, Mar del Plata, Argentina, 2000.

VOORHEES, E. M., "Query Expansion Using Lexical-Semasttic Relations", in *proceedings of ACMKHGIR'94*,pp. 61-69, 1994.

VOORHEES, E. M. "On test collections for adaptative information retrieval", *Information Processing and Management: an International Journal,* 44(6), 2008.

W3C, "Resource Description Framework (RDF)", disponível em http://www.w3.org/RDF/, 2004, acessado em março de 2012.

WANG, P., HU, J., ZENG, H.-JUN, CHEN, L., CHEN, Z., "Improving Text Classification by Using Encyclopedia Knowledge", *Seventh IEEE International Conference on Data Mining ICDM 2007*, *8*, 332-341. Ieee, 2007. Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4470257.

WANG, Q., WOO, H. L., QUEK, C. L., YANG, Y., LIU, M., "Using the *Facebook* group as a learning management system: An exploratory study", *British Journal of Educational Technology*, 43(3), no-no. doi:10.1111/j.1467-8535.2011.01195.x, 2011.

WAZLAWICK, R.S., "Metodologia de Pesquisa para Ciência da Computação", Editora Elsevier, 2009.

WENGER, E., "Communities of practice, a brief introducion", *Communities*, 1-5. Wenger, Etienne. 2011.

WOOLDRIDGE, M.; JENNINGS, N., "Intelligent agents – Theory and practice." *Knowledge Engineering Review*, v.10, n.2, 1995.

WOOLDRIDGE, M. J., "An Introduction to Multiagent Systems." John Wiley & Sons - Chichester, England,2002.

XU, JINXI, ANDW. BRUCE CROFT, “Query expansion using local and global document analysis”. *In Proc. SIGIR*, pp. 4–11. ACM Press. 194, 522, 533, 1996.

YATES, R. B., NETO, B. R., “Modern Information Retrieval”. 1 ed, Addison Wesley, 1999.

YIN, R. K., “Estudo de Caso: Planejamento e Métodos”., 3 edição, Porto Alegre, Ed. Bookman, 2005.

ZAVAGLIA, C., OLIVEIRA, L., NUNES, M., “Estrutura Ontológica e Unidades Lexicais: uma aplicação computacional no domínio da Ecologia”. 200.169.53.89, 1575-1584, 2007. Retrieved from http://200.169.53.89/download/CD_congressos/2007/SBC2007/pdf/arq0162.pdf.

ZIMMER, M., “Web Search Studies: Multidisciplinary Perspectives on Web Search Engines”. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International Handbook of Internet Research*, 1994, (pp.507-521). Springer, Netherlands. 2010. Retrieved from http://www.springerlink.com/index/10.1007/978-1-4020-9789-8.

ZHUHADAR, L., NASRAOUI, O., “Semantic Information Retrieval for Personalized E-Learning”. *IEEE Conf. on Tools with Artificial Intelligence (ICTAI 2008)* (pp.364-368). IEEE Press, New York, 2008.

Apêndice I – Instruções Segundo Estudo de Caso

Este anexo mostra as atividades necessárias para a participação dos alunos no estudo de caso.

I.1. Instruções para a participação do estudo de caso "CCSA - Collaborative Context-Search Agent"

1) Todos os alunos deverão associar-se ao grupo **FSI 2011.2 (UNIRIO)**, criado na rede social *Facebook* (http://pt-br.facebook.com/groups/241589899235871/), conforme figura a seguir.

2) O ambiente será de aprendizagem colaborativa e participativa. As discussões realizadas nesse grupo seguirão a abordagem construtivista de ensino. Em outras palavras, os alunos deverão ler, postar e responder tópicos entre si, de acordo com o tema proposto.

3) O tema proposto é composto pelo conteúdo da primeira parte da disciplina até o dia 2 de outubro de 2011, ou seja, de "Introdução a Sistemas de Informação" *até* "Business Intelligence/Business Analytics". Os tópicos podem ser encontrados no moodle em *Moodle* > FSI2011.2 (Fundamentos de Sistemas de Informação 2011.2).
4) O horário para a realização das atividades é livre.
5) A pesquisa a documentos *Web* deve ser feita por intermédio de uma postagem no grupo, que deve seguir a sintaxe ***?? termos da consulta ??*****.** A figura abaixo ilustra o procedimento de busca.

6) Os termos da consulta devem obedecer aos mesmos critérios adotados pelo usuário em atividades de busca cotidianas em outros mecanismos de pesquisa, como o Bing, Google etc.
7) O protótipo deve ser o **único** mecanismo de aquisição de conhecimento para os participantes desta pesquisa. *Sites* de pesquisa *Web*, como Google e Bing não poderão ser utilizados. Para realizar uma consulta, os alunos devem seguir os passos do item 5 e aguardar a resposta do agente com o *link* para o *site* do protótipo.
8) Os alunos **devem** combinar os termos (coluna //Sugestões) que julgarem relevantes e clicar no botão Buscar. É desejável reexecutar a busca selecionando mais ou menos termos quantas vezes forem necessárias para testar o protótipo, até que se deseje avaliar os resultados. Nesse momento, todos os resultados da busca devem ser avaliados numa escala de 5 estrelas, onde uma estrela representa nenhuma relevância e 5 estrelas representa relevância máxima.

Relevância	Descrição
	Documento totalmente relevante.
	Documento relevante, suficiente para suprir a minha necessidade de informação.
	Documento parcialmente relevante, ou que contenha um *link* para outro documento de classificação 4 ou 5 estrelas.
	Pouco relevante e que não supra a minha necessidade de informação, mas com alguma relação ao tema pesquisado.
	Documento absolutamente irrelevante e fora do contexto.

A figura abaixo ilustra a interface do protótipo de busca.

9) O objetivo do protótipo é o auxílio à recuperação de documentos *Web* é encontrar aqueles que sirvam como fonte de informação para que os alunos possam entender ou participar mais da discussão.
10) O **questionário** será divulgado no moodle no último dia da pesquisa. A participação é obrigatória. Outros avisos e orientações gerais também poderão ser publicados no moodle a qualquer momento. Dúvidas também poderão ser encaminhadas por e-mail a qualquer momento.

Muito Obrigado,

Eduardo Fritzen (fritzen@gmail.com)

Apêndice II – Questionário Segundo Estudo de Caso

Neste anexo serão listadas todas as perguntas do questionário respondido pelos alunos do estudo de caso. O objetivo deste questionário foi avaliar qualitativamente a utilização do protótipo *Collaborative Context Search*. As perguntas foram agrupadas em duas seções: identificação do perfil do aluno e experimento. Os participantes desta pesquisa foram informados que os dados obtidos nesta pesquisa poderão ser divulgados exclusivamente pelo pesquisador e seu orientador na literatura especializada, incluindo possivelmente revistas e eventos científicos da área, com a premissa de garantia de sigilo sobre suas identidades.

II.1. Identificação e Perfil

As perguntas marcadas com * são de preenchimento obrigatório.

- Nome:
- E-mail:
- No geral, qual grupo de sugestões trouxe melhores resultados para sua necessidade de informação?
- Idade (somente números)
- Quantas vezes acessa o *Facebook* por semana?
- Há quanto tempo você faz parte do *Facebook*?
- Quanto tempo gasta em média em cada acesso?
- Quantas postagens costuma fazer por semana?

- Quantos comentários costuma fazer por semana?
- Quanto tempo por semana passa pesquisando na *Web* material da disciplina?
- Quanto tempo por semana passa pesquisando coisas em geral?
- Quanto tempo por semana fica interagindo com os amigos pela *Web*?
- Como prefere trabalhar?
- Com relação às possíveis dificuldades encontradas ao pesquisar conteúdos na *Web* para a disciplina? (Marque até 3 opções):

II.2. Experimento

- Os resultados exibidos nos grupos "Grupo A", "Grupo B" e "Grupo C" representam diferentes visões em relação ao tema pesquisado?
- Nenhuma combinação de termos entre os grupos "A", "B" e "C" apresentaram resultados relevantes?
- Os termos dos grupos "B" e "C" abordaram assuntos novos, ou seja, expansões das discussões no grupo?
- Com que número de termos você considera ter obtido melhores resultados?
- O protótipo ajudou a encontrar palavras relevantes do contexto?
- Em média, a resposta para as solicitações feitas ao agente está de acordo com suas expectativas de tempo?
- Acessar a busca em uma nova aba atrapalhou a sua interação com o grupo?
- As tarefas de busca e avaliação dos resultados atrapalharam o andamento das discussões?

- Os resultados que obteve no protótipo ajudaram a participar mais das discussões?
- Os termos do grupo "C" estavam exclusivamente relacionados as discussões no grupo, ou seja, nenhum termo relacionado a um assunto diferente foi abordado?
- As sugestões de termos apresentados nos grupos "A", "B" e "C" estavam relacionadas ao mesmo assunto?
- Obteve mais resultados de páginas da *Wikipédia* quando escolheu uma sugestão de termo do grupo "A"?
- Durante a espera pela resposta do Agente, participou das discussões?
- Em consultas distintas, as palavras apresentadas no "Grupo C" são sempre muito semelhantes?
- Teve necessidade de alterar a expressão de busca e desistiu de encontrar o que procurava?
- As sugestões de termos apresentados nos grupos "A", "B" e "C" estavam relacionadas a sua necessidade de informação?
- Ver o que seus colegas de classe estavam buscando influenciou sua busca também?
- O fato de poder ver o que meus colegas estavam buscando possibilitou a intervenção e ajuda a estes colegas de classe?
- Com o passar do tempo, a qualidade das sugestões (melhorou - piorou):
- Deixe neste espaço outros comentários e sugestões sobre o protótipo.

Apêndice III – Comportamentos do Jade Utilizados

A arquitetura proposta faz uso do *framework Jade*. Este trabalho adotou o *Jade (Java Agent DEvelopment Framework)* por ser um *framework open source*[77] baseado na linguagem de programação *Java* e compatível com as especificações da FIPA, além de permitir o desenvolvimento de sistemas multiagentes e oferecer uma infraestrutura completa para o envio e recebimento de mensagens. Nesse *framework*, cada agente está associado a uma *thread*, que controla a execução de um ou mais comportamentos. Os comportamentos são ações que um agente pode desempenhar dentro de uma modelagem de sistema multiagente. Todas as tarefas dos agentes são executadas por meio de comportamentos, ou *behaviours*.

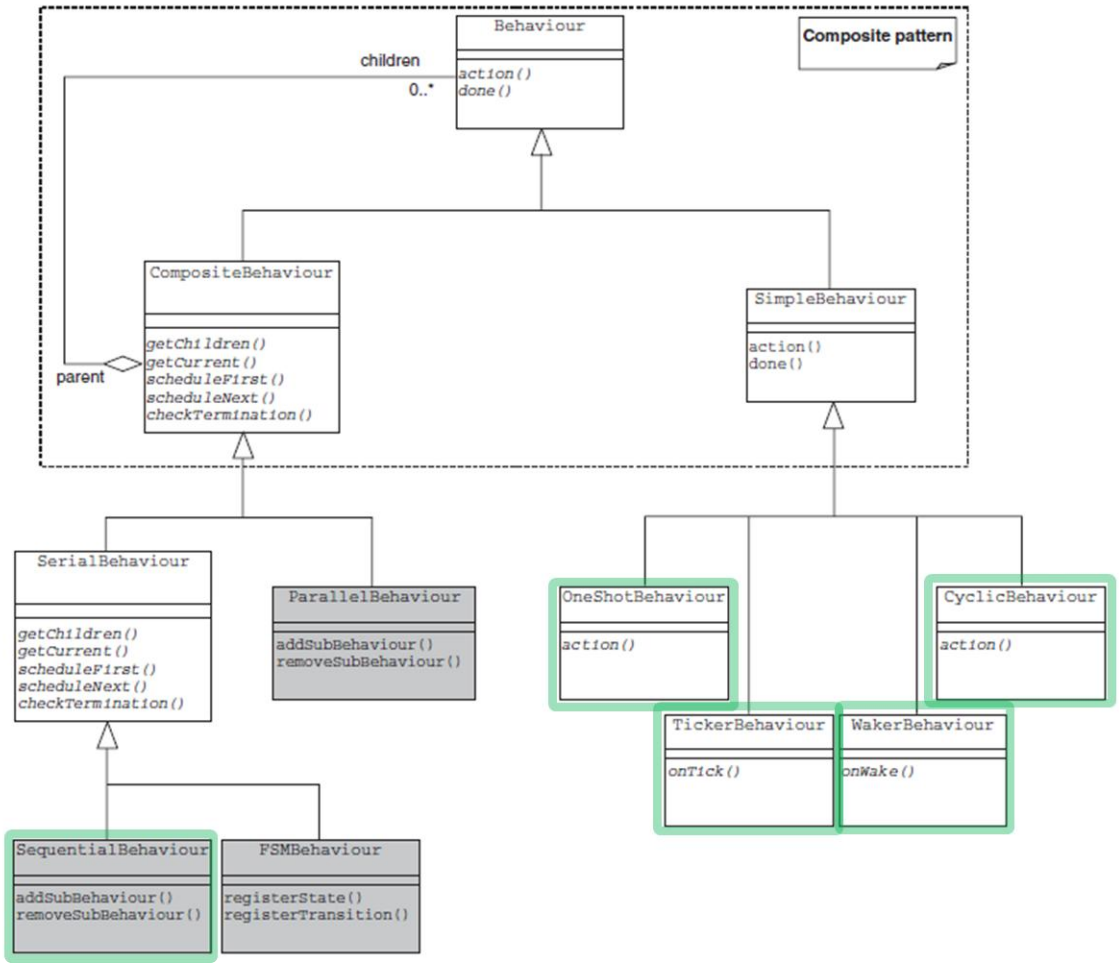
Um escalonador, presente na classe *Agent*, gerencia automaticamente o agendamento da execução dos comportamentos, ou seja, a chamada ao método de execução de um comportamento ("*public void action();"*) é realizada de forma transparente e não requer intervenção de programação para tanto. Na política de escalonamento, os comportamentos são executados de maneira circular e seguem o modelo não preemptivo[78], para todos os comportamentos aptos a execução em uma fila (BELLIFEMINE, 2007). Em outras palavras, dado um agente, somente um comportamento deste pode ser executado num momento do tempo e o método "*public*

[77] *Jade* é desenvolvido pela empresa de telecomunicações Italia Lab, e está disponível gratuitamente, sob licença LGPL (Lesser General Public License Version 2.0).
[78] Não existe distinção de prioridades, nem distribuição de tempo de execução entre os comportamentos.

void action();" deste agente deve executar até o fim para que outro comportamento seja executado pelo escalonador.

O *Jade*, assim como o *Java* é uma linguagem multiplataforma. É possível instalar a plataforma *Jade* em máquinas distribuídas na rede e, com isso, fazer com que os agentes dispostos em diferentes pontos da rede se comuniquem de forma simplificada. Por fim, o *Jade* possui aplicação gráfica para execução, controle, depuração e manutenção dos agentes (BELLIFEMINE, 2007).

A figura abaixo, extraída de (BELLIFEMINE, 2007, p. 92), apresenta a hierarquia de classes dos comportamentos definidos no *Jade* e destaca aqueles utilizados na arquitetura.



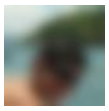
Comportamentos do *Jade*

- *SequentialBehaviour* – Usado para compor tarefas complexas. As subtarefas que compõem este tipo de comportamento são executadas sequencialmente, até que a última tarefa filha seja executada.
- *CyclicBehaviour* – Comportamento que deve ser executado continuamente, como, por exemplo, um comportamento que escute o ambiente para o envio e recebimento de mensagens.
- *TickerBehaviour* – Usado quando o comportamento deve executar tarefas periódicas.
- *OneShorBehaviour* – Comportamento que é executado uma única vez.
- *WakerBehaviour* – Este comportamento executa após aguardar um tempo limite determinado.

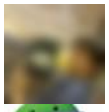
Apêndice IV – Mensagens Durante a Dinâmica

Esta seção traz um trecho das conversas entre os alunos e sua interação com o agente de busca durante a participação no segundo estudo de caso.

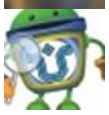
Alex Gomes
Qual o papel de um arquiteto de TI



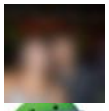
Guilherme Castro o arquiteto de TI diferentemente dos arquitetos corporativos é um profissional que alem de ter que ter um vasto conhecimento em tecnologia , tem que pensar nas melhores tecnologias disponiveis para a empresa em paralelo com os recursos que ela possui.
4 de Dezembro de 2011 às 12:39 · Curtir



Helio Santos ??Plano de TI??
4 de Dezembro de 2011 às 13:44 · Curtir

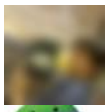


Unibot Agent Oi Helio Santos! Aqui está o link para a sua busca "Plano de TI".
http://unirio.noip.org/ CollaborativeContextSearch/ ?id=250161708378690
(via Context Search Agent)
4 de Dezembro de 2011 às 14:01 · Curtir

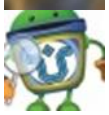


Marcos Araújo ??arquiteto de ti??
4 de Dezembro de 2011 às 19:52 · Curtir

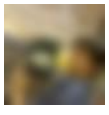
Unibot Agent Oi Marcos Araújo! Aqui está o link para a sua busca "arquiteto de ti".
http://unirio.no-ip.org/ CollaborativeContextSearch/ ?id=250360508358810
(via Context Search Agent)
4 de Dezembro de 2011 às 20:39 · Curtir



Helio Santos ??Arquiteto de TI??
6 de Dezembro de 2011 às 20:44 · Curtir

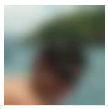


Unibot Agent Oi Helio Santos! Aqui está o link para a sua busca "Arquiteto de TI".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=251536658241195
(via Context Search Agent)
6 de Dezembro de 2011 às 21:19 · Curtir

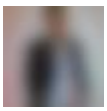


Helio Santos O papel de um arquiteto de TI é assegurar o tratamento adequado aos requisitos da empresa ou negócio o qual ele atende, conectando múltiplos pontos de vista sobre os sistemas, construir uma fundação sólida para suportar eventuais mudanças e adaptar o sistema ao tempo e recursos disponíveis ao negócio
6 de Dezembro de 2011 às 21:26 · Curtir

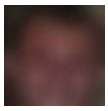
Marcelo Costa Curiosidade: Guilherme Castro, conseguiu formular esta resposta sem apoio de ferramentas de busca?
6 de Dezembro de 2011 às 21:49 · Curtir



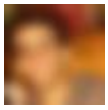
Guilherme Castro consegui , pois eu tinha pesquisado um pouco sobre o que um arquiteto de TI faz na pratica, um pouco antes de começar o trabalho.
7 de Dezembro de 2011 às 08:29 · Curtir



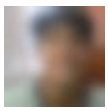
Alex Gomes
Quando é vantajoso um ERP???



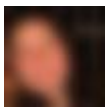
Marcelo Carvalho Entendo que a primeira coisa a ser feita é a avaliação do custo X benefício da implementação de um sistema ERP. Lembro do Sean comentando algo sobre empresas que tentaram implantar este tipo de sistema e acabaram falindo. A aquisição deste tipo de sistema requer uma análise personalizada de cada empresa. Não se deve considerar, portanto, que se o sistema foi extremamente eficiente em um determinado empreendimento ele o vá ser em qualquer empreendimento em que se tente implantar.
1 de Dezembro de 2011 às 18:37 · Curtir



Fábio Rocha Caso a empresa já possua uma boa estabilidade financeira, já que como o Claudio falou, os custos são elevados e criam uma dependência da empresa com o fornecedor do ERP, e a empresa esteja necessitando otimizar o fluxo de informação, da tomada de decisão e eliminar interfaces manuais, acredito que seja vantajoso. A implementação desse sistema requer um estudo detalhado antes da aquisição.
1 de Dezembro de 2011 às 19:36 · Curtir



Cláudio Reis Também devemos levar em conta certos aspectos, como a necessidade de adaptação do ERP para empresa, uma vez q uma empresa com atividades específicas não encontrarão um ERP que atenda a todas as suas necessidades, outro aspecto a ser visto, é sobre a utilização de um ERP de código aberto ou fechado, pois o aberto apesar de possibilitar a mudança do sistema, não tem suporte, enquanto o erp de código fechado, causa uma dependência muito grande da empresa contratada.
1 de Dezembro de 2011 às 20:57 · Curtir

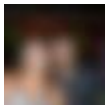


Aline Teixeira Para avaliar se é vantajoso ou não, acredito que se levarmos em conta quais os objetivos da empresa, poderemos encontrar algum pró ou contra à implementação do ERP, por exemplo se pode ser prejudicial a dependência da empresa com o fornecedor do ERP, questão que o Daniel já havia levantado.
1 de Dezembro de 2011 às 21:56 · Curtir

Unibot Agent Além do alto custo dos Sistemas integrados de gestão empresarial (ERP em inglês), a dependência do fornecedor é, em geral, uma outra grande desvantagem.
2 de Dezembro de 2011 às 03:07 · Curtir

Gabriel Fonseca ?? ERP ??
2 de Dezembro de 2011 às 12:36 · Curtir

Unibot Agent Oi Gabriel Fonseca! Aqui está o link para a sua busca "ERP".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=248873075174220
(via Context Search Agent)
2 de Dezembro de 2011 às 12:38 · Curtir

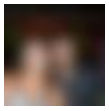


Marcos Araújo ?? ERP ??
2 de Dezembro de 2011 às 13:21 · Curtir

Unibot Agent Oi Marcos Araújo! Aqui está o link para a sua busca "ERP".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=248893268505534
(via Context Search Agent)
2 de Dezembro de 2011 às 13:23 · Curtir

Gabriel Fonseca Pelo o que eu entendi, para determinar se o ERP será vantojoso, primeiro deve-se avaliar se a empresa tem sistemas em um quantidade considerável, para que então se tenha o que alinhar. Ou seja, para uma empresa com poucos sistemas, não é vantajoso.
Outro ponto, seria avaliar as necessidades da empresa e implementar um ERP de acordo com suas características.
2 de Dezembro de 2011 às 13:31 · Curtir

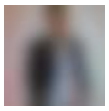
Gabriel Fonseca O ERP, por si só, pode ser muito vantajoso para otimizar o fluxo de informação, eliminar redundâncias dos sistemas e integrar os processos da empresa como um todo. Mas exige também, um estudo aprofundado das funções da empresa e exige dos usuários um certo conhecimento para manuseá-lo corretamente.
2 de Dezembro de 2011 às 13:32 · Curtir



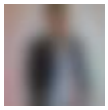
Marcos Araújo como já foi dito, devido ao alto custo do ERP e as vezes a por não se adequar as características de uma pequena empresa, acho que não deve ser vantajosa sua implementação em uma pequena empresa, a não ser que muito bem planejada. No entanto em uma grande empresa com uma grande quantidade de fluxo de informações acho que seria vantajosa a implementação desse sistema.
2 de Dezembro de 2011 às 20:56 · Curtir

Marcelo Costa Ficaram discutindo a questão do custo e a dependência de fornecedor... Alguns falaram sobre o fluxo de dados através das diversas atividades e setores da empresa... Entretanto, ainda faltou considerar os benefícios de uma automação centralizada e a dificuldade de adaptação de certos componentes...
6 de Dezembro de 2011 às 20:39 · Curtir

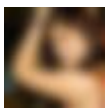
Marcelo Costa Alex Gomes, o que vc diz?
6 de Dezembro de 2011 às 20:40 · Curtir



Alex Gomes Os riscos são grandes, porém os benefícios de um ERP bem adaptado são muitos. As decisões podem ser tomadas de forma mais rápida e eficiente na empresa, otimizando o setor de estoque e venda entre outros. Um sistema bem integrado também pode proteger dados sensíveis, consolidando múltiplos sistemas de segurança em uma única estrutura. Entretanto esses sistemas são caros e nem sempre se adaptam perfeitamente as necessidades da empresa, muitos desses sistemas vem com o código fechado, impedindo certos tipos de adaptação. Empresas de pequeno porte podem encontrar dificuldade por não ter um capital tão grande para contratar um sistema totalmente adaptado, essas empresas podem tentar se adequar ao sistema ou podem optar por um outro sistema não tão eficiente mas com seu código aberto a mudanças adaptando ele as especificações da empresa.
6 de Dezembro de 2011 às 23:06 · Curtir



Alex Gomes
Quando é vantajoso um ERP???

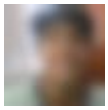


Marcele Brandão Bem basicamente quem a sua utilização a finalidade, se for individual é ferramenta pessoal e se for usada pra auxiliar a evolução de um grupo é a outra opção.
4 de Dezembro de 2011 às 15:49 · Curtir

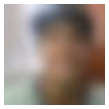
Marcelo Costa Mais respostas?
6 de Dezembro de 2011 às 21:48 · Curtir

Bruno Silveira Estas "ferramentas de trabalho em grupo" seriam os sistemas colaborativos?
6 de Dezembro de 2011 às 22:01 · Curtir

Helio Santos Não necessariamente, uma ferramenta de trabalho em grupo, por exemplo um web chat por si só não é um sistema colaborativo
6 de Dezembro de 2011 às 22:08 · Curtir



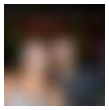
Cláudio Reis
Qual o momento certo para se fazer um plano de negócios?



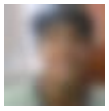
Cláudio Reis ??plano de negócios??
30 de Novembro de 2011 às 18:05 · Curtir

Unibot Agent Oi Cláudio Reis! Desculpe-me, mas ainda não tenho conhecimento suficiente para gerar a recomendação. Responderei sua solicitação assim que possível.
(via Context Search Agent)
30 de Novembro de 2011 às 20:01 · Curtir

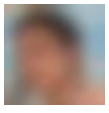
Victor Cunha acho que depende da complexidade do seu negócio e se existe tal necessidade ou não, cai naquela questão que falamos em sala, o dono da barraquinha de pipoca necessita de um plano de negócios? Na minha opnião não, é possível gerenciar o negócio sem precisar elaborar um plano de negócios mas pra ele pode ser necessário sim, ele pode melhorar o negócio dele se fizer? talvez, fazer um plano de negocios por fazer não significa resolver os problemas encontrados
30 de Novembro de 2011 às 20:43 · Curtir



Marcos Araújo bom eu acho que mesmo sem querer ao abrir um negócio mesmo que seja uma barraquinha de pipoca, o dono da barraquinha de pipoca deve ter levado certas questões em consideração como a localização da sua barraquinha, se seria na frente de uma escola ou de um cinema, qual seu publico alvo, se ele vai ter capital para se manter no inicio, acho que mesmo que incompleto a ideia de plano de negócios está enraizada na alma de qualquer negócio
30 de Novembro de 2011 às 21:23 · Curtir



Cláudio Reis Um plano de negócios tem como um dos objetivos resolver os problemas encontrados, não tem como você pensar em futuro, em desenvolvimento da empresa, sem antes pensar nos problemas que ela vem sofrendo ou que podem acontecer. A consideração de localização da barraquinha está mais pra "feeling" do que para plano de negócios.
Acredito que o plano de negócios deveria ser visto como uma das primeiras ações a serem tomadas antes de começar um negócio, podendo variar de complexidade de a acordo com o tamanho da empreitada. Lembrem-se, muitas vezes os planos de negócio são criados e apresentados antes mesmo da criação da empresa.
30 de Novembro de 2011 às 21:47 · Curtir

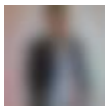


Leandro Soares Eu acho que o ideal seria fazer o plano de negócios antes da empresa entrar em ação. A partir daí, ela poderá conhecer seus futuros concorrentes, no que ela precisa de mais investimentos... O que não quer dizer que uma empresa que já está ativa há um tempo não possa começar a fazer o seu plano de negócios se não possui. Como falaram, depende da necessidade de cada uma.
30 de Novembro de 2011 às 22:05 · Curtir

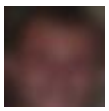
Alberto Moraes Penso que a elaboração de um Plano de negócios é essencial para o empreendedor, não somente para a busca de recursos, mas principalmente, como forma de planejar de forma mais eficiente suas ideias, antes de entrar de

cabeça no mercado, que é muito competitivo.
30 de Novembro de 2011 às 22:27 · Curtir

Pedro Duarte eu acredito que o plano de negócios tem que ser uma das primeiras coisas a serem feitas,pois ela ajuda a planejar suas idéias e como coloca-las em pratica e tambem tem a questão de problemas futuros
que podem ser muito bem precentido no plano de negocios e assim teriamos logo uma solução para o problema, pois sempre é bom ter um plano 'B'. Mesmo q seja um simples negocio o empreendedor leva em considerações algumas coisas q possam ser colocadas como plano de negócios, por isso acho que o plano de negócios deve fazer parte de qualquer empresa, embora algumas possam crescer sem tem elaborado um plano de negócios!
1 de Dezembro de 2011 às 09:15 · Curtir

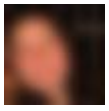


Alex Gomes Eu concordo com a ideia apresentada pelo Claudio, qualquer negócio seja ele do menor ao maior é necessário se fazer um estudo antes. Não que seja necessário contratar alguém para isso, ou fazer um plano escrito mas uma análise da situação é sempre necessário.
1 de Dezembro de 2011 às 12:47 · Curtir

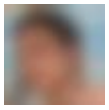


Marcelo Carvalho Eu entendo ser necessário que o PN seja elaborado antes de se iniciar um empreendimento porque é um documento que resulta de um estudo sobre o ambiente do negócio que se quer iniciar. Acredito também que, depois de iniciado o empreendimento, possam ser feitas algumas modificações.
1 de Dezembro de 2011 às 13:26 · Curtir

Mônica Abreu Como o Marcelo Carvalho disse, o plano de negócios é para ser feito, de preferencia, antes de iniciar um negócio em si. Dessa forma, é possivel prever certos acontecimentos e fazer uma análise do mercado no futuro para que a pessoa que quer começar o negócio esteja preparada.
1 de Dezembro de 2011 às 16:43 · Curtir

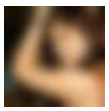


Aline Teixeira Com relação ao momento, o ideal seria fazer o PN antes de iniciar o negócio, mas isso não quer dizer que ele não possa ser feito depois que o negócio foi iniciado. Se deve ser feito ou não, penso que é escolha de quem vai abrir o negócio. Na minha visão ele sempre deveria ser feito, pois acredito que com ele podemos entender em que contexto o negócio estará inserido e até antecipar soluções para problemas futuros, apesar de existirem negócios bem sucedidos que nunca fizeram um PN; mas acredito que mesmo esses que não fizeram um PN, levam algumas questões dele em consideração.
1 de Dezembro de 2011 às 21:19 · Curtir

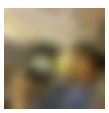


Leandro Soares Acho que a mesma coisa vale para o plano de TI. Inclusive, a empresa vai ter um melhor resultado se fizer o plano de TI junto com o plano de negócios,conseguindo uma melhor integração.
1 de Dezembro de 2011 às 21:30 · Curtir

Unibot Agent Oi Cláudio Reis! Aqui está o link para a sua busca "plano de negócios".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=247911235270404
(via Context Search Agent)
2 de Dezembro de 2011 às 00:52 · Curtir



Marcele Brandão ??plano de negócios??
2 de Dezembro de 2011 às 09:21 · Curtir

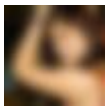


Helio Santos Concordo com o Marcelo Carvalho, um planejamento de negócios deve ser realizado antes do início do empreendimento, pois se iniciarmos o empreendimento com um estudo sobre o ambiente de negócio e outros diversos estudos, torna-se mais fácil conseguir vantagens comerciais
2 de Dezembro de 2011 às 11:09 · Curtir

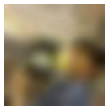
Gabriel Fonseca Também concordo com o Marcelo Carvalho, pois o plano de negócios é fundamental para conseguir sócios e investidores. A idéia do negócio é muito valiosa, e mesmo sem capital para iniciação , o plano de negócios pode ser

feito e a partir daí surgir o negócio.
2 de Dezembro de 2011 às 12:18 · Curtir

Unibot Agent Oi Marcele Brandão! Aqui está o link para a sua busca "plano de negócios".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=248758155185712
(via Context Search Agent)
2 de Dezembro de 2011 às 12:19 · Curtir



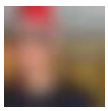
Marcele Brandão Percebi que em uma discussão inicial houve a questão sobre quem necessita fazer um PN e quem não precisa. Antes de existir um PN, com esse nome, os negociantes já tinham algum tipo de planejamento para que o seu negócio fosse a frente uma previsão de estoque, um antecipamento da quantidade de vendas, depois quando houve um aumento demasiado da concorrencia é que o PN foi mais necessário, foi preciso esse nivel de detalhamento nas previsoes de problemas pra se ganhar o mercado. Logo, vejo que estar em um ambiente muito concorrido é um dos grandes motivos de se fazer um PN.
4 de Dezembro de 2011 às 15:23 · Curtir



Helio Santos Creio que o momento seja o momento da criação do negócio, pois a administração e operação do mesmo torna-se mais fácil quando você tem uma visão do negócio como um todo, ao invés de lidar com situação por situação na medida em que elas aparecem
6 de Dezembro de 2011 às 20:45 · Curtir

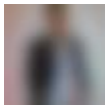
Marcelo Costa A discussão ficou no que cada um acha... Me pergunto se procuraram buscar sobre o histórico de plano de negócios, bem como informações sobre empresas como SEBRAE que apoiam a implantação de empresas (o que elas pensam sobre o assunto)... Seria interessante investigarem outras fontes para não ser apenas um "acho" de vocês, mas um relato de estudos e experiências... De qualquer modo, isto será melhor discutido em AEA...
6 de Dezembro de 2011 às 21:47 · Curtir

Cláudio Reis Eu li alguma coisas no sebrae, e no site da caixa, sobre investimentos em micro e pequenas empresas, logo, o planejamento do negócio, independente do tamanho, é necessário, pois isso diminui a margem de erro do negócio, porém, mesmo q a recomendação seja fazer um plano de negócios, existem diversos casos aonde o empreendedor simplesmente abre uma loja, um quiosque e começa o trabalho.
6 de Dezembro de 2011 às 21:59 · Curtir · 1



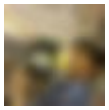
José Macedo
?? Sistemas de informação??

Unibot Agent Oi José Macedo! Aqui está o link para a sua busca "Sistemas de informação".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=250557405005787
(via Context Search Agent)
5 de Dezembro de 2011 às 05:54 · Curtir



Alex Gomes ?? Sistemas de informação??
5 de Dezembro de 2011 às 16:48 · Curtir

Unibot Agent Oi Alex Gomes! Aqui está o link para a sua busca "Sistemas de informação".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=250838251644369
(via Context Search Agent)
5 de Dezembro de 2011 às 16:50 · Curtir

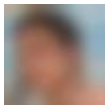


Helio Santos Sistemas de Informação são todos os sistemas que tem como intenção captar, analisar, transmitir e transformar dados que representem informações aos usuários.
6 de Dezembro de 2011 às 20:39 · Curtir

Marcelo Costa Lembrem-se que tem outra discussão sobre o mesmo assunto...
6 de Dezembro de 2011 às 21:59 · Curtir

Cláudio Reis
O que deve ser avaliado quando uma empresa resolve expandir seus negócios para outros países?

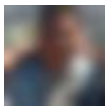
Victor Cunha as leis desse país, cultura também pois determinados negócios podem ter rejeição devido a cultura de tal país
4 de Dezembro de 2011 às 13:26 · Curtir



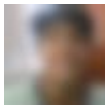
Leandro Soares Deve avaliar também como anda o mercado nesse país, os concorrentes, as oportunidades...
4 de Dezembro de 2011 às 14:37 · Curtir

Mônica Abreu Além disso, questão de beneficios, impostos, mão de obra, preço de tecnologia,...
4 de Dezembro de 2011 às 21:01 · Curtir

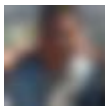
Cláudio Reis mas caso o possível negócio possa ter rejeição nesse país, deve-se abandonar a expansão? redefinir a abordagem do negócio tbm é possível, por exemplo, uma país como a china que tem mais de 1,3 bilhão de pessoas não pode ser tratado como um simples país de expansão, talvez reinventar um negócio, especificamente para esse país seja mais vantajoso do que não expandir por causa da rejeição...
4 de Dezembro de 2011 às 21:08 · Curtir



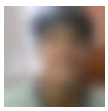
Victor Cunha o que quis dizer foi que se vc for uma empresa que vende carne bovina vai se instalar na india onde nao se consome devido a religião? é pedir pra falir =] por isso que tem que se avaliar a questão cultural de cada país
4 de Dezembro de 2011 às 21:11 · Curtir



Cláudio Reis eu compreendi, era exatametne sobre isso que tava falando, seu exemplo ainda foi muito bom, porque por exemplo, existe mcdonalds na índia, haveria rejeição se eles n tivessem mudado o tipo de produto que eles fornecem a esse país, no caso, são hamburgers sem carne bovina por causa da cultura do país e sem carne de porco, por causa da grande quantidade de muçulmanos no país
4 de Dezembro de 2011 às 21:16 · Curtir

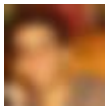


Victor Cunha mas dai nao é reinventar o negocio e sim seguir para outros ramos afim de atender a demanda, ela vendendo carne bovina ela nao reinventou fez outro tipo de carne revolucionaria e começou a vender lá, a partir do estudo feito ela comprou outro tipo de animal para a produçao de alguma carne que pudesse ser vendida dai concordo com o que tu falou anteriormente, de redefinir a abordagem porem de fato existirão casos onde nao havera tal possibilidade
4 de Dezembro de 2011 às 21:23 · Curtir



Cláudio Reis o entendimento de reinventar não foi o esperado, o que eu quis dizer foi mudar o produto, a abordagem, a propaganda, ou qualquer elemento que necessitasse de mudança para ser aceito no pais alvo, além do mais, o critério de entendimento de reinventar poderia ser isso que vc falou, de criar um tipo de carne revolucionário, ou apenas mudar os hamburgers de redondos para quadrados, seguir outros ramos seria o mcdonalds passar a vender carros
4 de Dezembro de 2011 às 21:27 · Curtir

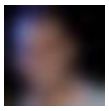
Marcelo Costa ;) assunto também discutido em AEA
6 de Dezembro de 2011 às 21:58 · Curtir



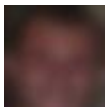
Fábio Rocha
O Facebook pode ser considerado um sistema colaborativo?

Alberto Moraes ??Sistema Colaborativo??
1 de Dezembro de 2011 às 19:04 · Curtir

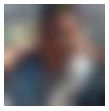
Unibot Agent Oi Alberto Moraes! Desculpe-me, mas ainda não tenho conhecimento suficiente para gerar a recomendação. Responderei sua solicitação assim que possível.
(via Context Search Agent)
1 de Dezembro de 2011 às 19:14 · Curtir



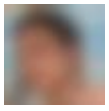
Rodrigo Souza Eu acho que o Facebook, assim como o Twitter, pode funcionar sim uma ferramenta de mídia colaborativa... mas é claro, desde que seja usado com esse objetivo.
1 de Dezembro de 2011 às 19:15 · Curtir



Marcelo Carvalho Um Sistema Colaborativo é baseado no Modelo 3C (Coordenação, Cooperação e Comunicação). Conforme a atividade que estamos desenvolvendo no momento, o facebook está funcionando com um ambiente colaborativo. O Pimentel nos contou que foi o primeiro a escrever um livro didático sobre Sistemas Colaborativos e afirmou que o Facebook é um Sistema Colaborativo, portanto, acho que ele tenha autoridade para atestar o que disse.
1 de Dezembro de 2011 às 19:18 · Curtir



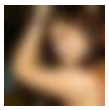
Victor Cunha Pra mim o facebook pode ser um sistema colaborativo mas a finalidade principal nao é de ser um sistema colaborativo e sim de ser um meio de comunicação e relacionamento entre pessoas, agora por exemplo a criação de um grupo para que membros possam interagir em relação a criação de um determinado projeto daí concordo que seja porque cai no conceito de pessoas com um mesmo objetivo ou com uma mesma tarefa se comunicando em tempo real ou não e diminuindo as barreiras impostas geograficamente
1 de Dezembro de 2011 às 21:16 · Curtir



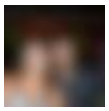
Leandro Soares Com isso, eu posso concluir que o Facebook ,a princípio, não é um sistema colaborativo mas possui ferramentas que podem ser consideradas colaborativas,no caso, os grupos.
1 de Dezembro de 2011 às 21:22 · Curtir

Unibot Agent Oi Alberto Moraes! Aqui está o link para a sua busca "Sistema Colaborativo".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=248443725217155
(via Context Search Agent)
2 de Dezembro de 2011 às 00:54 · Curtir

Unibot Agent Existe diferença entre Software colaborativo e Sistema Colaborativo?
2 de Dezembro de 2011 às 03:33 · Curtir



Marcele Brandão ??Software colaborativo??
2 de Dezembro de 2011 às 09:28 · Curtir



Marcos Araújo ??software colaborativo??
2 de Dezembro de 2011 às 09:28 · Curtir

Unibot Agent Oi Marcele Brandão! Aqui está o link para a sua busca "Software colaborativo".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=248760631852131
(via Context Search Agent)

2 de Dezembro de 2011 às 12:20 · Curtir

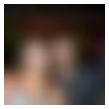
Unibot Agent Oi Marcos Araújo! Aqui está o link para a sua busca "software colaborativo".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=248760855185442
(via Context Search Agent)
2 de Dezembro de 2011 às 12:20 · Curtir

Gabriel Fonseca ?? Software Colaborativo ??
2 de Dezembro de 2011 às 12:21 · Curtir

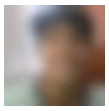
Unibot Agent Oi Gabriel Fonseca! Aqui está o link para a sua busca "Software Colaborativo".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=248866368508224
(via Context Search Agent)
2 de Dezembro de 2011 às 12:23 · Curtir

Gabriel Fonseca Quanto ao facebook, eu acho sim que é um sistema colaborativo, até porque teve um aluno na nossa turma, por exemplo, que criou o facebook somente para realização desta tarefa, ou seja, ele enxerga isso como um meio de contextualização acadêmica.
2 de Dezembro de 2011 às 12:23 · Curtir

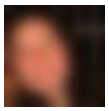
Gabriel Fonseca Pelo o que eu li, não há diferença.
"Skip Ellis, professor universitário norte-americano especializado em Groupware, define a ferramenta como um "sistema baseado em computador que auxilia grupos de pessoas envolvidas em tarefas comuns (ou objetivos) e que provê interface para um ambiente
compartilhado"
2 de Dezembro de 2011 às 12:27 · Curtir



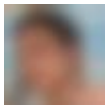
Marcos Araújo Concordo com o Gabriel Fonseca, não creio que exista diferença entre sistema e software colaborativos: "Sistemas Colaborativos são ferramentas de software utilizadas em redes de computadores para facilitar a execução de trabalhos em grupos."
2 de Dezembro de 2011 às 20:39 · Curtir



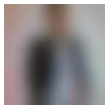
Cláudio Reis Um sistema colaborativo não cairia na mesma discussão que as ferramentas de produtividade pessoal? ou seja, o que define se é ou um não é, é o motivo pelo qual foi criado, o facebook não foi criado com a intenção de ser um sistema colaborativo, mas isso não impede que ele seja usado de tal forma.
2 de Dezembro de 2011 às 21:23 · Curtir



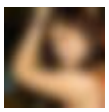
Aline Teixeira Concordo com o Pedro, acho que cairia na mesma discussão. E se sistemas colaborativos são ferramentas de software, e quanto ao exemplo da pizza, aquele não era um sistema colaborativo?
2 de Dezembro de 2011 às 21:30 · Curtir



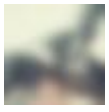
Leandro Soares Sim,aquele exemplo da pizza também era um exemplo de sistema colaborativo.Havia cooperação,comunicação e coordenação.
2 de Dezembro de 2011 às 21:36 via celular · Curtir



Alex Gomes sistema colaborativo é todo sistema que contém cooperação, comunicação e coordenação. O facebook pode ou não ser um sistema colaborativo, depende da maneira que ele vai ser usado, mas ele tem o necessário para ser um sistema colaborativo.
3 de Dezembro de 2011 às 01:08 · Curtir

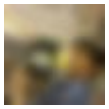


Marcele Brandão Não foi criado para issso, mas dependendo da maneira como for usado, por exemplo, nós aqui, pode ser sim um sistema colaborativo e ter cooperação, comunicação e coordenação.
4 de Dezembro de 2011 às 15:36 · Curtir



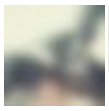
Mônica Abreu Também concordo com o Cláudio Reis, não é porque uma ou duas pessoas o utilizam como sistema colaborativo que ele irá se tornar um. Acho que o que conta mais é para que ele foi criado.

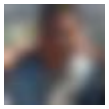
4 de Dezembro de 2011 às 21:03 · Curtir

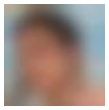

Helio Santos O facebook não deve ser considerado um sistema colaborativo, o intuito do idealizador do facebook era somente de uma rede social intra-colegial, mesmo que este ideal tenha se modificado. Um sistema colaborativo tem sua origem na necessidade da solução de um problema.
6 de Dezembro de 2011 às 20:37 · Curtir

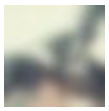

Marcelo Costa Aqui é importante notar um sistema colaborativo como um sistema que serve para colaboração... Sim, observem estes sistemas em relação ao modelo 3C, isto pode ajudar a categorizar... Também reparem em sistemas na empresa e fora dela (ou seja, podemos observar um sistema colaborativo mesmo que a colaboração não seja para atingir um objetivo de trabalho)...
6 de Dezembro de 2011 às 21:56 · Curtir


Marcelo Costa Para FSI a discussão está ok, mas podem seguir com o tema em outra disciplina: Sistemas Colaborativos.
6 de Dezembro de 2011 às 21:57 · Curtir

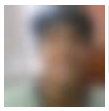

Mônica Abreu
Como é utilizado o sistema de gestão de conhecimento nas empresas?


Victor Cunha palestras seria um exemplo? por exemplo digamos que seja uma empresa de consultoria de mercado, um funcionário bastante experiente daria palestras comentando como analisar tendências, identificar oportunidades baseado na sua experiência
4 de Dezembro de 2011 às 13:30 · Curtir


Leandro Soares Treinamento dos funcionários também entraria nesse contexto?
4 de Dezembro de 2011 às 14:36 · Curtir

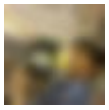

Mônica Abreu Assim, pelo que eu acho, eles também utilizam a questão de colocar as coisas na nuvem, como por exemplo, a wikipedia. As empresas estão utilizando suas próprias versões e cada funcionario compartilha alguma coisa.
4 de Dezembro de 2011 às 21:00 · Curtir


Marcelo Costa Mais respostas? (Lembrando que temos um capítulo no livro do Stair & Reynolds sobre isto... e temos a Internet via a ferramenta de busca do Unibot Agent...)
6 de Dezembro de 2011 às 21:51 · Curtir

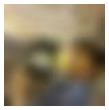

Cláudio Reis
Para que serve um Dicionário de Dados ?

Bruno Silveira Pra especificar os possíveis valores que podem ser atribuidos aos domínios de um determinado atributo
4 de Dezembro de 2011 às 12:53 · Curtir

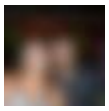
Bruno Silveira Esses atributos são aqueles que obtemos ao traçar as relações , baseadas no DER após serem aplicados os 7 passos do mapeamento
4 de Dezembro de 2011 às 12:57 · Curtir


Helio Santos ??Dicionario de Dados??
4 de Dezembro de 2011 às 13:32 · Curtir


Unibot Agent Oi Helio Santos! Aqui está o link para a sua busca "Dicionario de Dados".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=250145675046960 (via Context Search Agent)
4 de Dezembro de 2011 às 13:34 · Curtir

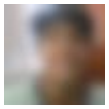


Helio Santos Um dicionario de dados serve para padronizar a notação dos dados que serão utilizados pelo sistema
4 de Dezembro de 2011 às 13:39 · Curtir



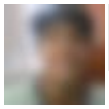
Marcos Araújo ??dicionario de dados??
4 de Dezembro de 2011 às 19:53 · Curtir

Unibot Agent Oi Marcos Araújo! Aqui está o link para a sua busca "dicionario de dados".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=250360928358768
(via Context Search Agent)
4 de Dezembro de 2011 às 20:41 · Curtir

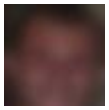


Cláudio Reis também serve pra ajudar na leitura dos modelos de dados
4 de Dezembro de 2011 às 20:54 · Curtir

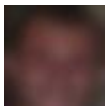
Marcelo Costa Mais respostas?
6 de Dezembro de 2011 às 21:50 · Curtir



Cláudio Reis
O que seria a visão da empresa?

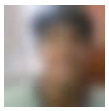


Marcelo Carvalho É a perspectiva da empresa a longo prazo. Onde e quando a empresa quer chegar.
4 de Dezembro de 2011 às 13:25 · Curtir

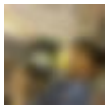


Marcelo Carvalho É a perspectiva da empresa a longo prazo. Onde e quando a empresa quer chegar.
4 de Dezembro de 2011 às 13:25 · Curtir

Marcelo Costa Mais respostas? Detalhamento?
6 de Dezembro de 2011 às 21:41 · Curtir

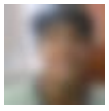


Cláudio Reis as visões seriam um conjunto de desejos, aspirações, os quais a empresa quer atingir, sem se preocupar em como fazer isso, talvez, de acordo com o andar do planejamento, a visão possa sofrer uma revisão, caso a visão inicial seja inatingível
6 de Dezembro de 2011 às 21:43 · Curtir



Helio Santos Seria o que a empresa espera no futuro, em um determinado espaço de tempo. A visão da empresa não reflete o negócio em si, nem objetivos quantitativos ou qulitativos, mostra apenas o que a empresa deseja ser vista
6 de Dezembro de 2011 às 21:46 · Curtir

Pedro Duarte
qual seria definição de objetivo e de meta??



Cláudio Reis as metas são objetivos a serem alcançados a longo prazo, como um objetivo principal da empresa, enquanto os objetivos são fins a serem atingidos em um prazo menor, o cumprimento desses objetivos são essenciais para o cumprimento da meta
4 de Dezembro de 2011 às 12:44 · Curtir

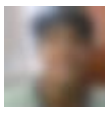
Marcelo Costa ;)
6 de Dezembro de 2011 às 21:41 · Curtir

Alberto Moraes
Qual a finalidade do Diagrama de Fluxo de Dados ??

Bruno Silveira ?? diagrama de fluxo de dados objetivos ??
4 de Dezembro de 2011 às 12:29 · Curtir

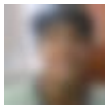
Unibot Agent Oi Bruno Silveira! Aqui está o link para a sua busca "diagrama de fluxo de dados objetivos".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=250111895050338
(via Context Search Agent)
4 de Dezembro de 2011 às 12:31 · Curtir

Bruno Silveira Se não me engano, é expor claramente o funcionamento das funções de um sistema
4 de Dezembro de 2011 às 12:31 · Curtir

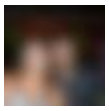


Cláudio Reis mostrar os processos internos de um sistema, no caso o msm sistema apresentado no diagrama de contexto
4 de Dezembro de 2011 às 12:41 · Curtir

Marcelo Costa Este tópico deveria ser discutido no youflow...
6 de Dezembro de 2011 às 21:40 · Curtir

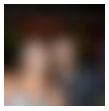


Cláudio Reis
Quais são os passos necessários para a criação de um plano estratégico ?

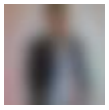


Marcos Araújo ?? plano estratégico ??
4 de Dezembro de 2011 às 05:30 · Curtir

Unibot Agent Oi Marcos Araújo! Aqui está o link para a sua busca "plano estratégico".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=249939155067612
(via Context Search Agent)
4 de Dezembro de 2011 às 05:32 · Curtir

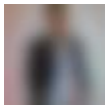


Marcos Araújo 1 - Execução de uma análise do ambiente
2 - Estabelecimento de uma diretriz organizacional
3 - Formulação de uma estratégia organizacional
4 - Implementação da estratégia organizacional
5 - Controle estratégico
4 de Dezembro de 2011 às 05:52 · Curtir



Alex Gomes ?? plano estratégico ??
4 de Dezembro de 2011 às 11:56 · Curtir

Unibot Agent Oi Alex Gomes! Aqui está o link para a sua busca "plano estratégico".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=250095618385299
(via Context Search Agent)
4 de Dezembro de 2011 às 11:56 · Curtir



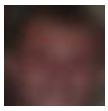
Alex Gomes Definição da missão corporativa.
Análise da situação.
Formulação de objetivos.
Formulação de estratégias.

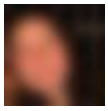
Implementação, Feedback e controle.
4 de Dezembro de 2011 às 12:05 · Curtir

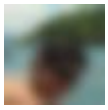

Marcelo Costa E aí, Cláudio Reis, era o que vc esperava como resposta?
6 de Dezembro de 2011 às 21:32 · Curtir

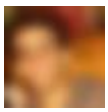

Marcelo Costa Que tal analisarem estas respostas em relação ao conjunto de itens que vimos/discutimos que estariam contemplados em um plano estratégico?
6 de Dezembro de 2011 às 21:33 · Curtir


Cláudio Reis eu não tenho certeza de quais são os itens contemplados, seriam, análise, planejamento, implementação e controle?
6 de Dezembro de 2011 às 21:39 · Curtir


Marcelo Carvalho
O que é Arquitetura de TI?


Aline Teixeira São o conjunto de padrões e serviços que estão relacionados com os componentes de TI.
3 de Dezembro de 2011 às 13:38 · Curtir


Guilherme Castro A arquitetura de TI tambem pode ser classificada em quatro tipos: Arquitetura estratégica,de projetos, de aplicação e tecnológica.
3 de Dezembro de 2011 às 15:03 · Curtir


Fábio Rocha ?? arquitetura de TI ??
3 de Dezembro de 2011 às 18:28 · Curtir


Unibot Agent Oi Fábio Rocha! Aqui está o link para a sua busca "arquitetura de TI".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=249674151760779
(via Context Search Agent)
3 de Dezembro de 2011 às 21:17 · Curtir

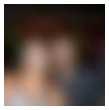

Cláudio Reis ??arquitetura tecnologia da informação??
3 de Dezembro de 2011 às 21:29 · Curtir


Unibot Agent Oi Cláudio Reis! Aqui está o link para a sua busca "arquitetura tecnologia da informação".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=249744365087091
(via Context Search Agent)
3 de Dezembro de 2011 às 21:33 · Curtir


Helio Santos ??Plano de TI??
3 de Dezembro de 2011 às 21:36 · Curtir

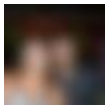

Unibot Agent Oi Helio Santos! Aqui está o link para a sua busca "Plano de TI".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=249746575086870
(via Context Search Agent)
3 de Dezembro de 2011 às 21:39 · Curtir


Cláudio Reis Trata-se da organização lógica de dados, dos serviços de TI e interfaces que serão disponibilizados a organização para atender aos seus objetivos estratégicos.
3 de Dezembro de 2011 às 22:31 · Curtir

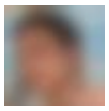


Marcos Araújo ?? arquitetura de ti ??
4 de Dezembro de 2011 às 05:30 · Curtir

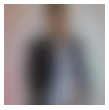
Unibot Agent Oi Marcos Araújo! Aqui está o link para a sua busca "arquitetura de ti".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=249939301734264
(via Context Search Agent)
4 de Dezembro de 2011 às 05:31 · Curtir



Marcos Araújo "A arquitetura de TI pode ser entendida como o conjunto de elementos constituintes da TI. A necessidade e o papel de TI na empresa constituem os fatores determinantes da arquitetura de TI"(Rodriguez e Ferrante). A arquitetura de TI engloba os seguintes componentes: Infra-estrutura de hardware, infra-estrutura de comunicação, sistemas de informação, sistemas de bancos de dados, metodologias
4 de Dezembro de 2011 às 06:19 · Curtir

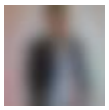


Leandro Soares A arquitetura de TI mostra como os objetivos da empresa serão alcançados através do uso da informação e tecnologia.
4 de Dezembro de 2011 às 11:16 · Curtir



Alex Gomes ?? arquitetura de ti ??
4 de Dezembro de 2011 às 11:56 · Curtir

Unibot Agent Oi Alex Gomes! Aqui está o link para a sua busca "arquitetura de ti".
http:// unirio.no-ip.org/ CollaborativeContextSearch/ ?id=250096038385257
(via Context Search Agent)
4 de Dezembro de 2011 às 11:59 · Curtir



Alex Gomes É a integração dos objetivos da empresa com a tecnologia.
4 de Dezembro de 2011 às 12:08 · Curtir

Marcelo Costa Interessante como vocês fizeram a ligação de arquitetura de TI com arquitetura de informação, com plano de TI e com planejamento estratégico... Isto é assunto de uma disciplina de mestrado e possivelmente deverá virar uma disciplina de Tópicos Avançados na graduação em breve... ;)
6 de Dezembro de 2011 às 21:35 · Curtir