UNIVERSIDADE DO SAGRADO CORAÇÃO

CHRISTIAN FREITAS

DESENVOLVIMENTO DE UM PROTÓTIPO PARA DETECÇÃO AUTOMÁTICA DE PERFIS UTILIZANDO MINERAÇÃO DE DADOS E ONTOLOGIA

BAURU

2015

UNIVERSIDADE DO SAGRADO CORAÇÃO

CHRISTIAN FREITAS

DESENVOLVIMENTO DE UM PROTÓTIPO PARA DETECÇÃO AUTOMÁTICA DE PERFIS UTILIZANDO MINERAÇÃO DE DADOS E ONTOLOGIA

Trabalho de Conclusão de Curso apresentado ao Centro de Ciencias Exatas e Sociais Aplicadas da Universidade do Sagrado Coração. Como parte dos requisitos para obtenção do titulo de Bacharel em Ciência da Computação, sob orientação do Prof. Me. Patrick Pedreira Silva.

BAURU

2015

Freitas, Christian

F8666d

Desenvolvimento de um protótipo para detecção automática de perfis utilizando mineração de dados e ontologia / Christian Freitas. -- 2015.

69f. : il.

Orientador: Prof. Me. Patrick Pedreira Silva.

Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade do Sagrado Coração – Bauru – SP.

1. Ontologia. 2. Inteligência Artificial. 3. Mineração de Dados. I. Silva, Patrick Pedreira. II. Título.

CHRISTIAN FREITAS

DESENVOLVIMENTO DE UM PROTÓTIPO PARA DETECÇÃO AUTOMÁTICA DE PERFIS UTILIZANDO MINERAÇÃO DE DADOS E ONTOLOGIA

Trabalho de Conclusão de Curso apresentado ao Centro de Ciências Exatas e Sociais Aplicadas da Universidade Sagrado Coração como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, sob a orientação do Me. Patrick Pedreira Silva

Banca examinadora:

_____

Prof. Me. Patrick Pedreira Silva
Universidade do Sagrado Coração

_____

Prof. Dr. Elvio Gilberto da Silva
Universidade do Sagrado Coração

_____

Prof. Me. Henrique Pacchioni Martins
Universidade do Sagrado Coração

Bauru, 02 de Dezembro de 2015

AGRADECIMENTOS

Agradeço primeiramente a minha família, pela força durante todo o curso.

Ao professor Me. Patrick Pedreira Silva pela dedicação e paciência para conclusão do trabalho de Iniciação Científica e o presente trabalho.

Aos docentes do curso que durante os anos contribuíram para a minha formação.

RESUMO

A quantidade de informações textuais geradas, principalmente na internet, fez com que pesquisadores procurassem cada vez mais formas de encontrar métodos de implementar ferramentas que possam processar e organizar tais informações de modo automático. Esse crescimento é dado pelo aumento do uso de serviços online, o que acarreta grande impacto no comportamento dos consumidores uma vez que eles têm acesso a uma grande variedade de produtos e informações. Nesse contexto, este projeto, consiste em desenvolver um software, utilizando a linguagem de programação JavaScript, que através das redes sociais seja capaz de agrupar informações e classifica-las, avaliando características linguísticas e estatísticas para a detecção automática de tópicos de um texto em português e para a geração de perfis, retornando ao usuário suas preferências de conteúdos na internet.

Palavras-chave: Inteligência Artificial. Mineração de Dados. Ontologias.

ABSTRACT

The amount of textual information generated, especially on the internet, has caused researchers to seek more and more ways to find methods to implement tools that can process and organize such information automatically. This growth is given by the increased use of online services, which entails major impact on consumer behavior since they have access to a wide variety of products and information. In this context, this project is to develop software using the JavaScript programming language, which through social network is able to collate information and ranks them by assessing linguistic characteristics and statistics for the automatic detection of topics of a text in Portuguese and generating profiles, returning to their user preferences in the internet content.

Keywords: Artificial Intelligence. Data mining. Ontologies.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Com o crescimento exponencial da quantidade de informações textuais geradas e compartilhadas, sobretudo com a internet, torna-se importante encontrar métodos e desenvolver ferramentas que possam processar e organizar tais informações de modo automático. Esse crescimento pode ser inclusive verificado com relação ao número de redes sociais e formas de compartilhar o conteúdo que mais lhe agrada, o que acarreta grande impacto no seu comportamento uma vez que eles têm acesso a uma grande variedade de informações.

Por um lado a facilidade de desfrutar de um serviço online gerou um novo mercado, por outro lado tornou-se mais difícil a escolha dos usuários por serviços e/ou produtos que melhor suprem suas necessidades e expectativas. Neste contexto, os sistemas de recomendação surgem como uma das principais soluções para o problema do grande volume de informações, pois permitem disponibilizar para os usuários sugestões personalizadas e automatizadas. Sistemas deste tipo podem ser usados em variadas aplicações desde a recomendação de livros, filmes até a recomendação de usuários com perfis semelhantes (em sites de relacionamento, por exemplo). O desafio da recomendação é estimar as boas indicações, de um serviço ou produto, por exemplo, para um usuário baseado no seu perfil, uma vez que as pessoas tendem a interagir com itens que são de seu interesse pessoal. O interesse de um usuário pode ser expresso de modo implícito, dentre outras formas, considerando informações textuais associadas a um perfil de rede social (itens compartilhados).

Para tais fins tem o processamento semântico automático por meio de ontologias vem se tornando cada vez mais comum nos últimos anos, já que elas fornecem meios de representar e utilizar o conhecimento de mundo. Em aplicações que envolvam o processamento de conteúdo textual, esse conhecimento de mundo pode significar, por exemplo, entender sobre o que versa um texto ou mesmo detectar tópicos que indiquem preferências de uma pessoa. As ontologias podem ser aplicadas para identificar em um texto-fonte seus tópicos principais a fim de subsidiar as mais diversas tarefas, entre elas: criação automática de sumários, classificação automática de documentos, geração automática de textos, definição de perfis de usuário, etc. Com o uso das ontologias é possível agregar valor à informação disponível, ou seja, associar um grau de utilidade às respostas geradas automaticamente por sistemas computadorizados de processamento de língua natural (PLN).

Diante deste potencial, este projeto propõe adotar o uso de uma ontologia, como uma das estratégias para a identificação de tópicos e como subsídio para o desenvolvimento de um software que utilize esse conhecimento para classificar um perfil de usuário.

A acepção utilizada neste trabalho é aquela que denota uma ontologia como uma especificação explícita de uma conceitualização acerca de um domínio. (GENESERETH; NILSSON, 1987; GRUBER, 1996).A conceitualização envolve, assim, a definição de uma coleção de conceitos que se assume existirem em um domínio, assim como os relacionamentos entre eles. (GENESERETH; NILSSON, 1987).

A conceitualização é expressa como uma representação do vocabulário terminológico da ontologia e das relações entre esses termos. (GRUBER, 1996). Neste contexto a ontologia a ser utilizada no projeto descreve uma taxonomia de palavras, consistindo de vocabulários de representação de conceitos, provendo termos potenciais para descreverem o conhecimento de um domínio, sendo utilizados como indicadores de informações relevantes. (GUARINO, 1994).

Um conjunto de conceitos será usado para formar a base de dados ontológica. Dessa forma, a ontologia será composta por uma coleção de conceitos relacionados e estes, por sua vez, serão representados por um conjunto de palavras. Os conceitos serão expressos através da língua natural, utilizando termos específicos, ou seja, palavras que, quando encontradas, indicam a presença do conceito. (LOH, 2001). Desta forma, poderão ser identificados através de técnicas que analisam o conteúdo textual dos documentos.

Com relação à sua estrutura, um conceito é composto de um identificador e de um conjunto de palavras que o descrevem, chamadas de descritores.

O identificador é uma palavra da língua natural que dá a ideia geral do conceito. (WIVES, 2004). Podem ser utilizados nomes de objetos, substantivos, verbos, etc., (“futebol”, “olimpíadas” e “jantar”, por exemplo). Os descritores do conceito são palavras que sinalizam a presença do conceito. Para definir os descritores podem-se utilizar o próprio identificador e outras palavras relacionadas. Por exemplo, o conceito Futebol poderia ser expresso por palavras identificadoras como: futebol, pênalti, gol, escanteio, impedimento, etc. (PEDREIRA-SILVA, 2006). As palavras descritoras podem estar relacionadas umas às outras de diferentes formas. As formas mais comuns de relacionamento são as sinonímias (relação de sentido entre dois vocábulos que têm significação muito próxima, permitindo que um seja escolhido pelo outro em alguns contextos, sem alterar o sentido literal da sentença como um todo), hiponímias (relação existente entre uma palavra de sentido mais específico e outra de

sentido mais genérico) e hiperonímias (relação estabelecida entre um vocábulo de sentido mais genérico e outro de sentido mais específico).

Além desses relacionamentos citados anteriormente, os descritores de conceitos também podem ser definidos por meio de variações morfológicas, tais como variações léxicas de gênero, número e grau, verbos e diferenças de grafia. (ex.: acrobata/acróbata)(AGIRRE et al., 2001; LOH, 2001; WIVES, 2004).

A definição de uma base de dados ontológica requer a coleta e descrição de conceitos que se pretende representar. Uma vez definida ela pode ser usada, então, para a identificação de tópicos em documentos a fim de determinar aqueles que sejam os mais importantes para caracterizar um perfil de usuário. A definição automática de perfis, abre a oportunidade para que serviços de recomendação possam fazer uso dessa informação.

A ferramenta a ser desenvolvida nesta pesquisa combina características linguísticas e estatísticas para a detecção automática de tópicos de um texto em português e mineração de dados para a geração de perfis. A hipótese norteadora deste trabalho é de que a informação semântica recuperada de uma ontologia, com base em informação textual associada aos usuários de redes sociais, permite que o sistema determine quais tópicos são relevantes para composição de um perfil que caracteriza um usuário. O sistema faz a identificação de tópicos pela contagem de conceitos, usando a ontologia do Yahoo Enriquecida adaptada de (PEDREIRA-SILVA, 2006) e, posteriormente combina essas informações objetivando a construção automática de um perfil.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Construir um software que faça a detecção automática de perfis de usuários utilizando teorias associadas à área de Processamento de Língua Natural e Inteligência Artificial.

2.2 OBJETIVOS ESPECÍFICOS

a) Enriquecer a ontologia desenvolvida Pedreira-Silva (2006) e integrá-la software;
b) Definir um modelo de perfis para ser utilizado pelo algoritmo de classificação;
c) Desenvolver algoritmo que faça a classificação de perfis dos usuários;
d) Avaliar o protótipo;
e) Quantificar resultados obtidos pelo software.

3 INTELIGÊNCIA ARTIFICAL

A palavra inteligência vem do latim inter (entre) e legere (escolher), tendo como significado “aquilo que dá o poder ao ser humano escolher entre uma coisa e outra, permitindo que realize uma tarefa de forma eficiente”. Já Artificial no latim significa artificiale, ou seja, algo não natural, produzido pelo homem. Por isso Inteligência Artificial é a inteligência criada pelo homem para dotar os computadores de algum tipo de habilidade semelhante ao raciocínio humano. (Fernandes, 2005).

Inteligência Artificial (IA) tem como objetivo o estudo e a modelagem da inteligência tratada como fenômeno, porém isso é extremamente complexo, por se tratar do resultado de milhões de anos de evolução da espécie humana. Uma das maiores que utilizam a IA é a robótica, onde cientistas tentam fazer com que as máquinas possam interagir com o meio assim como os seres humanos.

De acordo com Russel e Norvig (2003), Inteligência Artificial é uma das ciências mais recentes do planeta, iniciada logo após a Segunda Guerra Mundial, em 1956. Nos dias de hoje a IA abrange uma enorme variedade de subcampos, desde a área de uso geral, como aprendizado e percepção, até tarefas específicas como jogos, demonstração de teoremas matemáticos, criação de poesia e diagnóstico de doenças.

Para Ganascia (1993 citado por FERNANDES, 2005), os principais modelos de Inteligência Artificial são os Algoritmos Genéticos, Programação Evolutiva, Lógica Fuzzy, Raciocínio Baseado em Casos, Programação Genética e Redes Neurais e Sistemas Baseados em Conhecimento.

Há diversas abordagens diferentes de Inteligência Artificial para solucionar uma grande variedade de problemas. Uma dessas abordagens é o aprendizado de máquina, uma das áreas mais relevantes dentro da IA. O aprendizado é qualquer processo no qual um sistema melhora seu desempenho através da experiência, como por exemplo, perceber o gosto de um usuário.

Na Ciência da Computação, a Inteligência Artificial é voltada para o desenvolvimento de sistemas de computadores inteligentes, isto é, sistemas que exibem caraterísticas que se associam com inteligência no comportamento humano, como por exemplo, a compreensão de uma linguagem ou resolução de um problema.

3.1 HISTÓRIA

O termo "Artificial Intelligence" foi criado por John McCarthy durante o famoso Workshop do DartmouthCollege em 1956. Aquele foi o primeiro encontro de cientistas oficialmente organizado para discutir aspectos de inteligência e sua implementação em máquinas. Naqueles dias havia um grande entusiasmo e alguns experimentos relativamente bem sucedidos, mesmo com o estágio primitivo dos computadores e linguagens de programação disponíveis. Um dos desenvolvimentos significativos dos anos que se seguiram foi o GPS (General Problem Solver), criado por Allen Newell e Herbert Simon para simular os métodos humanos de resolução de problemas. Porém antes de John McCarthy a Inteligência Articial já existia. (SANTOS, 2006).

Nos anos 40 em meados da segunda Guerra Mundial a necessidade do avanço tecnológico para fornecer mais instrumentos para o combate bélico não somente a inteligência artifical, mas a computação num geral teve um avanço incalculável. Havia um ramo de pesquisas interessado na realização da representação das células nervosas do ser humano no computador, uma vez que o cérebro é formado de neurônios e é ele que realiza o processamento das informações do corpo. Esta linha de pesquisas motivou o desenvolvimento de uma formalização matemática para o neurônio, estabelecendo o neurônio formal. Esta formalização permitiu a realização de diversas concepções matemáticas sobre a forma de aprendizado dos neurônios, ou seja, como os neurônios armazenam informações. (SANTOS, 2006).

Na década de 50 iniciou-se o estudo, na linha de pesquisa psicológica, da utilização da lógica de estratégia para finalidades matemáticas, como a prova de teoremas. Iniciou-se, também, a modelagem através de regras de produção, regras estas baseadas na lógica de predicados. A introdução da programação através de comandos de lógica de predicados proporcionou um grande avanço para a programação de sistemas que utilizassem esquemas de raciocínio. Daí foi possível o aperfeiçoamento do que já existia: jogos, aplicações matemáticas e simuladores. Mas, passando à história da linha biológica, essa década foi de grande sucesso dada à implementação do primeiro simulador de redes neurais artificiais e do primeiro neurocomputador. A partir do modelo matemático de MacCulloc e Pitts (1943) e da teoria de aprendizado de Donald Hebb (1949), foi possível nessa década a união desses conhecimentos no modelo de rede neural artificial chamado Perceptron. (SANTOS, 2006).

Em 60, prosseguiram os desenvolvimentos de conceitos relativos às redes neurais artificiais com o aprimoramento do modelo Perceptron e o surgimento de uma variante, o

Adaline. Ambos utilizavam as mesmas ideias de rede, porém a lógica de aprendizado os diferenciava. Para a linha psicológica essa década foi a descoberta da Inteligência Artificial. É nesse momento que começa os pensamentos sobre ser possível realizar tarefas humanas, tais como o pensamento e a compreensão da linguagem, através do computador. (SANTOS, 2006).

Na década de 70 na linha biológica. Houve pesquisadores que, por outros caminhos, chegaram a novas concepções de redes neurais artificiais. Estas concepções analisavam o aprendizado de informações como sendo fruto de uma união das potencialidades de redes de neurônios interagindo entre si. Nasciam as redes neurais representadas na forma de mapas cerebrais, onde não havia o aprendizado de um neurônio, mas de toda uma rede, através do compartilhamento de recursos. Já na linha psicológica, estudos mais aprofundados demonstraram o óbvio: que não seria possível a representação numa máquina dos estados mentais humanos responsáveis pelo pensamento. (SANTOS, 2006).

Nos anos 80 as redes neurais artificiais tiveram seu reconhecimento recuperado através do físico John Hopfield, que em 1982 provou ser possível a simulação de um sistema físico através de um modelo matemático baseado na teoria das redes neurais. Assim, em 1986, uma equipe de especialistas das mais diversas áreas reuniram-se para validar as pesquisas em torno das redes neurais, possibilitando a volta da pesquisa nesta linha. Uma das formas de recuperação do prestígio das redes neurais foi a proposta de um modelo, chamado Backpropagation, que ampliava o potencial do Perceptron de modo a permitir a superação das limitações do modelo primitivo. Houve também o interesse de trabalho conjunto com outras áreas, tais como interfaces inteligentes, sistemas de apoio à decisão, controle de robôs, etc. (SANTOS, 2006).

Por volta dos anos 90, nessa década, as redes neurais tiveram uma explosão exponencial de aplicações e desenvolvimento de modelos. São centenas de propostas de novos ou aperfeiçoamento de modelos a cada ano, tal o interesse pela área. A partir daí, consolidam-se as redes neurais como parte integrante do estudo da Inteligência Artificial propriamente dita. Reconhece-se, também, que os paradigmas biológico e psicológico são complementares e necessários para sistemas mais evoluídos. Desta forma, começam nesta década a serem construídos os chamados Sistemas Híbridos. Estes sistemas são a união das concepções das duas linhas de pesquisa, permitindo a construção de grandes sistemas que pretendem abranger uma forma mais completa de representação do comportamento humano. Sendo este o ideal da própria Inteligência Artificial. (SANTOS, 2006).

4 PROCESSAMENTO DE LÍNGUA NATURAL

O Processamento de Linguagem Natural é uma subárea da Inteligência Artificial responsável pela construção de mecanismos que permitam a interpretação, em nível computacional, de sentenças muito utilizadas na linguagem do ser humano (JUNIOR, 1999).

4.1 APLICAÇÕES PRÁTICAS

Há várias tarefas comuns atualmente que incluem em maior ou menor grau capacidades linguísticas as quais o computador, estando num carro, celular, televisão ou qualquer outro aparelho, pode desempenhar.

Através de programas apropriados e softwares com capacidades linguísticas que podem ser um poderoso auxiliar tarefas do cotidiano. Vários exemplos desses softwares e que são exemplificados por Isaías (1995):

a) Escrita e produção de um texto: No campo da ajuda à redação incluem-se programas que detectam erros (detectores), sugerem alternativas (corretores), disponibilizam recursos – como tesauros e dicionários (monolíngues, bilíngues, de sinônimos) – e fornecem ajuda de gramática.

b) Tradução: A tradução é uma das atividades que envolvem mais conhecimento linguístico, visto que codifica a informação presente no texto de uma língua num texto de outra língua. Não é de estranhar que tenha sido a primeira área em que se trabalhou em PLN.

c) Aprendizagem e ensino: sistemas que respondem a perguntas ou juntam, de uma forma "apresentável", informação sobre um dado tema são, ou podem ser, auxiliares preciosos no processo de aprendizagem, delegando no computador uma parte da atividade de ensinar.

d) Sistemas interativos: dar comandos a uma máquina através da fala é já possível, embora de forma rudimentar, em vários sistemas sofisticados, como carros, aviões, ou casas assim como na interação com programas no trabalho ou em casa, como editores de texto e organizadores pessoais.

e) Indexação: identificação de livros e outros materiais de coleção, assim como de textos e revistas eletrônicas, é preciso uma incessante atividade de indexação, independentemente de estarmos a falar de bibliotecas digitais. (ISAÍAS, 1995).

f) Segurança e identificação: sistemas baseados em características únicas do falante podem ser usados para permitir que um sistema apenas reaja à "voz do dono".

5 ONTOLOGIAS

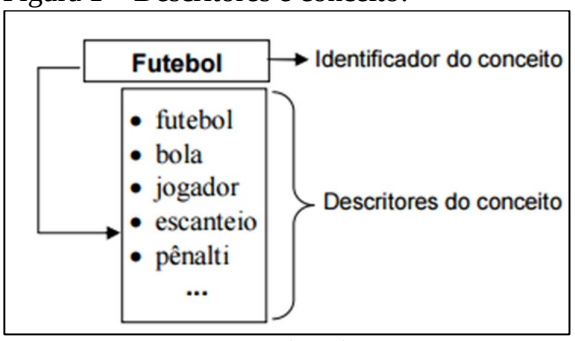
O termo ontologia origina-se da filosofia referindo-se a um ramo que lida com a organização e a natureza do ser associados a questões como: “O que é um ser?” e “Quais são as características comuns de todos os seres?”. (MAEDCHE, 2002). Entretanto esse termo foi adotado também pela área de Gestão de Conhecimento para se referir a conceitos e termos que podem ser usados para descrever alguma área do conhecimento ou construir uma representação dessas.

Neste contexto, Gomez-Pérez (1999) define uma ontologia como um conjunto de termos ordenados hierarquicamente para descrever um domínio que pode ser usado como um esqueleto para uma base de conhecimentos. Deste modo uma ontologia deve possuir um conjunto de termos organizados como uma hierarquia associada de classes com suas relações, restrições, axiomas e termologia associada, provendo uma estrutura básica na qual se pode construir uma base de conhecimento.

A ontologia fornece um conjunto de conceitos e termos para descrever um determinado domínio, enquanto a base de conhecimento usa esses termos para descrever uma determinada realidade, podendo ser usada por sistemas computacionais.

Neste sentido uma utiliza-se um conjunto de conceitos ara formar uma base de dados ontológica. Dessa forma, uma ontologia é formada por uma coleção de conceitos relacionados e estes, por sua vez, são representados por um conjunto de palavras. Os conceitos podem ser expressos através da língua natural, utilizando termos específicos, ou seja, palavras que, quando encontradas, indicam a presença do conceito. (LOH, 2001). Desta forma, podem ser identificados através de técnicas que analisam o conteúdo textual dos documentos. Com relação à sua estrutura, um conceito é composto de um identificador e de um conjunto de palavras que o descrevem (Figura 1) chamadas de descritores.

Figura 1 – Descritores e conceito.

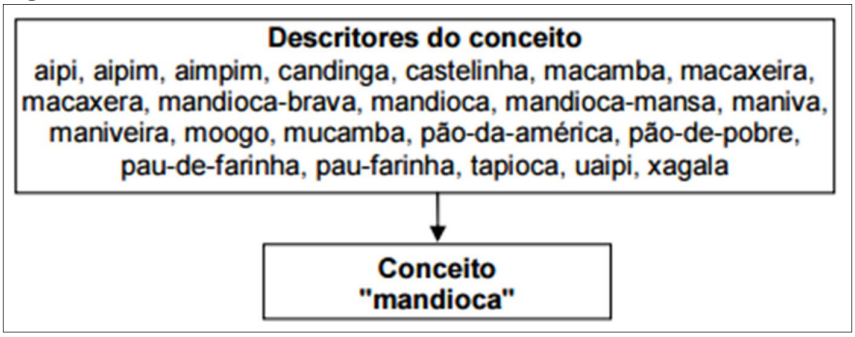


Fonte: PEDREIRA-SILVA (2006).

O identificador é uma palavra da língua natural que dá a ideia geral do conceito (WIVES, 2004). Podem ser utilizados nomes de objetos, substantivos, verbos, etc., (“basquete”, “doença” e “jantar”, por exemplo). Os descritores do conceito são palavras que sinalizam a presença do conceito. Para definir os descritores podem-se utilizar o próprio identificador e outras palavras relacionadas. Por exemplo, o conceito Futebol poderia ser expresso por palavras identificadoras como: futebol, pênalti, gol, escanteio, impedimento, etc. As palavras descritoras podem estar relacionadas umas às outras de diferentes formas. As formas mais comuns de relacionamento são as sinonímias1, hiponímias2 e hiperonímias3 (TIUN et al., 2001). A Figura 2 é um exemplo fictício de um conceito denominado mandioca cujos descritores foram definidos com base na relação de sinonímia.

[1] Relação de sentido entre dois vocábulos que têm significação muito próxima, permitindo que um seja escolhido pelo outro em alguns contextos, sem alterar o sentido literal da sentença como um todo.
[2] Relação existente entre uma palavra de sentido mais específico e outra de sentido mais genérico.
[3] Relação estabelecida entre um vocábulo de sentido mais genérico e outro de sentido mais específico.

Figura 2 – Descritores no formado de sinônimos.



Fonte: PEDREIRA-SILVA (2006).

Além desses relacionamentos citados anteriormente, os descritores de conceitos também podem ser definidos por meio de variações morfológicas, tais como variações léxicas de gênero, número e grau, verbos e diferenças de grafia (ex.: acrobata/acróbata) (AGIRRE et al., 2001; LOH, 2001; WIVES, 2004).

De acordo com Novello (2002), o uso de ontologias torna possível definir uma infraestrutura para integrar sistemas inteligentes no nível do conhecimento. O autor cita ainda algumas vantagens no uso de ontologias:

- Colaboração, o que possibilita uma interdisciplinaridade;
- Interoperação, o que facilita a integração da informação, em especial, em sistemas distribuídos;
- Informação, podendo assim usar como fonte de consulta e referência, o domínio;
- Modelagem, em que as ontologias são representadas por blocos estruturados, podendo ser reusáveis na modelagem de sistemas no nível de conhecimento
- Permitir o reuso de conceitos em domínios.

Dependendo de suas características as ontologias podem ser classificadas de diferentes modos. (MAEDCHE, 2002; GOMEZ-PÉREZ, 1999). Maeche (2002) faz a classificação baseada numa característica chave das ontologias, que utiliza a conceitualização como o principal critério para classificação. A seguir, são apresentados os quatro tipos propostos de classificação de uma ontologia:

Ontologias de alto-nível descrevem conceitos muito gerais como espaço, tempo, evento, etc. Esses conceitos tipicamente são independentes de um problema particular ou domínio. Sendo assim, é bem razoável ter-se uma ontologia de alto-nível compartilhada por grandes comunidades de usuários.

a) Ontologias de domínio: Descrevem o vocabulário relacionado a um domínio genérico, através da especialização de conceitos introduzidos nas ontologias de alto-nível. São exemplos de ontologia de domínio ontologias de veículos, documentos, etc.

b) Ontologias de tarefa: Descrevem um vocabulário relacionado a uma tarefa ou atividade genérica, através da especialização de conceitos introduzidos nas ontologias de alto-nível.

c) Ontologias de aplicação: São as ontologias mais específicas por serem utilizadas dentro das aplicações. Esse tipo de ontologia especializa conceito tanto das ontologias de domínio, como também das de tarefas.

No âmbito do Processamento Automático das Línguas Naturais e da Inteligência Artificial, outro pesquisador (VOSSEN, 1998a), aponta para uma classificação de dois tipos: as chamadas ontologias linguísticas e as ontologias conceituais. (VOSSEN, 1998a)

As ontologias linguísticas caracterizam-se por armazenar apenas conceitos lexicalizados (em uma determinada língua), isto é, conceitos expressos por uma ou mais palavras de uma língua. Sob esse ponto de vista, uma ontologia é um inventário dos sentidos de uma dada língua, ou seja, é um inventário somente daqueles conceitos compartilhados por uma comunidade linguística. Nesse sentido, uma ontologia linguística do holandês, por exemplo, não armazenaria o conceito “container”, já que este não é lexicalizado nessa língua. (VOSSEN, 1998a)

As ontologias conceituais, por sua vez, caracterizam-se pelo armazenamento de conceitos para os quais não há lexicalizações, ou seja, não há unidades lexicais que os representem, por exemplo: os conceitos “coisa parcialmente temporal” e “partes do corpo humano”. (VOSSEN, 1998a; PALMER, 2001)

6 MINERAÇÃO DE DADOS

6.1 INTRODUÇÃO

Resumidamente mineração de dados trata-se de extrair ou minerar conhecimento de grandes volumes de dados.

“Mineração de dados é a exploração e a análise, por meio automático ou semi automático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos”. (BERRY; LINOFF, 1997, p.5).

Os principais objetivos da mineração de dados são descobrir relacionamentos entre dados e fornecer subsídios para que possa ser feita uma previsão de tendências futuras, detecção de perfil e outros.

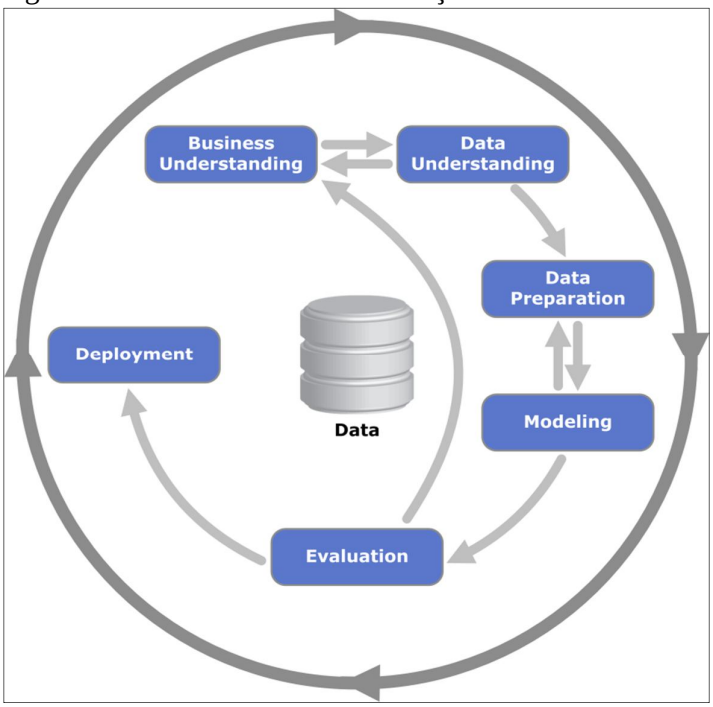
Os resultados obtidos com a mineração de dados podem ser usados no gerenciamento de informação, processamento de pedidos de informação, tomada de decisão, controle de processo e muitas outras aplicações.

6.2 FASES DA MINERAÇÃO

O ciclo em um processo de mineração de dados é constituido por 6 fases. A seqüência de fases não é obrigatória, ocorrendo a transição para diferentes fases, dependendo do resultado de cada fase, e que etapa particular de cada fase precisa ser executada em seguida. As setas indicam as mais importantes e mais freqüentes dependências entre as fases.(AMORIM, 2006).

A Figura 3 tem como objetivo exemplificar cada fase dentro de um ciclo.

Figura 3 – Ciclo das fases da mineração.



Fonte: AMORIM, 2006.

Segue abaixo uma síntese das etapas supracitadas.

a) Entendimento do Negócio (Business Understanding)

Nessa etapa tem o foco no entendimento do negócio, qual o objetivo do colhimento dos dados. Essa fase é importante, pois através do entendimento dela se pode tomar estratégias distintas para conseguir chegar ao objetivo. (AMORIM, 2006).

b) Seleção dos dados (Data Understanding)

É uma fase muito importante, pois é nela que serão decididos quais são os conjuntos de dados que serão relevantes para atingir os objetivos.

Essa fase começa com uma coleta de dados, e com procedimentos e atividades visando a familiarização com os dados, para identificar possíveis problemas de qualidade, ou detectar subconjuntos interessantes para formar hipóteses. (AMORIM, 2006).

c) Limpeza dos Dados (Data Preparation)

Antes que os algoritmos comecem a utilizar os dados é feito uma limpeza, pois os dados podem conter “sujeiras”, entenda sujeira como dados que possam afetar o resultado, afim de conseguir um resultado mais próximo do objetivo.

Este processo de limpeza dos dados geralmente envolve filtrar, combinar e preencher valores vazios.(CAMILO; SILVA, 2009)

d) Modelagem dos dados (Modeling)

Nessa fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para se obter valores otimizados. Geralmente, existem várias técnicas para o mesmo tipo de problema de mineração. Algumas técnicas possuem requerimentos específicos na forma dos dados. Consequentemente, voltar para a etapa de preparação de dados é frequentemente necessário.

A maioria das técnicas de mineração de dados são baseadas em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação e clusterização.(AMORIM, 2006).

e) Avaliação do processo (Evaluation)

A avaliação tem como objetivo garantir que o modelo gerado atente as expectativas, em outras palavras que o modelo atingiu o objeto pré-determinado e determinando que a Mineração foi um sucesso.

Nesta etapa é necessária a participação de especialistas nos dados, conhecedores do negócio e tomadores de decisão. Diversas ferramentas gráficas são utilizadas para a visualização e análise dos resultados. (CAMILO; SILVA, 2009).

f) Execução (Deployment)

É a etapa final onde consiste implantação do projeto ou entrega dos dados dependendo exclusivamente do tipo de projeto solicitado.

6.3 TÉCNICAS

A mineração de dados é desenvolvida para diferentes fins, por conta disso podem extrair diferentes tipos de conhecimento sendo assim decidir já no inicio do processo qual o tipo de conhecimento que o algoritmo deve extrair.

Existem 5 (cinco) técnicas gerais de mineração de dados que englobam todas as outras formas de apresentação e permitem uma visão mais global e apropriada ao assunto. (AMORIM, 2006). São elas:

a) Classificação

A classificação é uma das técnicas mais usadas pois se aproxima-se muito de atividade humanas na ajuda de compreender o ambiente.

Euriditionhome (2004, apud AMORIM,2006) diz que em um processo de mineração de dados, a classificação está especificamente voltada à atribuição de uma das classes pré-definidas pelo analista a novos fatos ou objetos submetidos à classificação. Essa técnica pode ser utilizada tanto para entender dados existentes quanto para prever como novos dados irão se comportar.

b) Estimativa

Estimar um valor, sendo ele um índice, temporal ou monetário, através dos dados que foram extraidos.

Suponha que se deseja determinar o gasto de famílias cariocas com lazer e que para isto se possua índices de gastos de famílias paulistanas com lazer, em função da faixa etária e padrão sóciocultural. Não se sabe exatamente quanto as famílias cariocas gastam com lazer mas se pode estimar baseando-se nos dados das famílias paulistanas. Certamente que esta estimativa pode levar a grandes erros, uma vez que Rio de Janeiro e São Paulo são cidades com geografias diferentes e que oferecem diferentes opções de lazer a seus habitantes. (AMORIM, 2006).

c) Desvio

A tarefa de desvio tem por objetivo descobrir um conjunto de valores que não seguem padrões definidos. Para estatarefa é necessário adotar antecipadamente.

Pode-se usar esta tarefa para identificar fraudes por meio de elementos que estao fora dos padrões ou são exceções as regras.

Essa técnica muito utilizada no comércio eletrônico afim de detectar fraudes de pagamento.

d) Análise de afinidades

Essa técnica tem por objetivo detectar padrões, ocorrências.

Um bom exemplo de análise de afinidades é o do carrinho de supermercado, do qual deseja-se conhecer quais os produtos que são normalmente são comprados em conjunto pelos consumidores.

Isto possibilita a otimização do layout interno dos supermercados e a realização de vendas dirigidas nas quais os itens são oferecidos já em conjuntos com preços menores.

e) Análise de agrupamentos

A análise de agrupamentos visa formar grupos mais homogêneos entre si. Pode ser estabelecido previamente um número de grupos a ser formado, ou se pode admitir ao algoritmo de agrupamento uma livre associação de unidades, de forma que a quantidade de grupos resultante seja conhecida somente ao final do processo.

É importante demonstrar que a diferença entre o agrupamento e classificação é que na classificação as classes são pré-definidas pelo pesquisador, enquanto que aqui não existe tal requisito.

7 PROCEDIMENTOS DE TESTES

Para o projeto foi utilizado o processo de prototipação, esse processo é a forma mais rápida e econômica de se definir e experimentar um projeto, esses dois motivos por si só já garantem sua importância, porém, ainda assim é comum vermos sistemas tomando forma antes de qualquer rascunho. (NASCIMENTO, 2013).

Segundo Nascimento (2013) o procedimento de prototipação possui varias classificações, sendo elas:

1. **Protótipos de Baixa Fidelidade:** Os protótipos de baixa fidelidade, também chamados de rascunhos ou sketches, são concebidos ainda na fase inicial, durante a concepção do sistema. Desenhados geralmente à mão utilizando lápis, borracha e papel, essas representações são feitas de maneira rápida e superficial, apenas margeando a ideia do projeto e definindo superficialmente sua interação com o usuário, não se preocupando ainda com elementos de layout, cores, disposições, etc.
2. **Protótipos de Média Fidelidade:** Conhecidos também por wireframes, esse protótipos são desenvolvidos na fase da arquitetura da informação. Utilizando lápis e papel ou softwares de prototipação, como o Balsamiq ou Axure, esses documentos apresentam a estrutura e o conteúdo da interface, definindo peso, relevância e relação dos elementos, formando o layout básico do projeto. Os protótipos de média fidelidade ainda não utilizam recursos gráficos avançados como cores ou fotografias.
3. **Protótipos de Alta Fidelidade:** Os mockups ou protótipos funcionais constituem a representação mais próxima do sistema a ser desenvolvido. Em alguns casos, é possível simular o fluxo completo das funcionalidades, permitindo a interação do usuário como se fosse o produto final. A aparência visual, as formas de navegação e interatividade já são concebidas e aplicadas aos protótipos de alta fidelidade.

Baseado no desenvolvimento de um protótipo com base no conhecimento dos requisitos iniciais para o sistema. O desenvolvimento é feito obedecendo à realização das diferentes etapas de análise de requisitos, o projeto, a codificação e os testes. Não necessariamente estas etapas devem ser realizadas de modo muito explícito ou formal. (PRESSMAN, 2005).

A definição de todos os requisitos necessários ao sistema pelo cliente ou usuário geralmente é uma tarefa muito difícil. É quase impossível prever como o sistema irá afetar o funcionamento das práticas de trabalho, como será a interação com outros sistemas e que operações dos usuários devem ser automatizadas. Mas para poder testar os requisitos de uma forma mais eficiente, seria necessária a utilização de um protótipo do sistema. (PRESSMAN, 2005).

Um protótipo é uma versão inicial de um sistema de software, que é utilizada para mostrar conceitos, experimentar opções de projeto e, em geral, para conhecer mais sobre os problemas e suas possíveis soluções. O desenvolvimento rápido de um protótipo é essencial para que os custos sejam controlados e os usuários possam fazer experiências com o protótipo no início do processo de software. (PRESSMAN, 2005).

Um protótipo de software apóia duas atividades do processo de engenharia de requisitos:

1. Levantamento de requisitos - Os protótipos de sistema permitem que os usuários realizem experiências para ver como o sistema apóia seu trabalho. Eles obtêm novas idéias para os requisitos e podem identificar pontos positivos e negativos do software. Eles podem, então, propor novos requisitos de sistema.
2. Validação de requisitos - O protótipo pode revelar erros e omissões nos requisitos propostos. Uma função descrita em uma especificação pode parecer útil e bem-definida. Contudo, quando essa função é utilizada com outras, os usuários muitas vezes acham que sua visão inicial era incorreta e incompleta. A especificação de sistema pode então ser modificada para refletir sua compreensão alterada dos requisitos.

Na maioria dos projetos, o primeiro sistema construído dificilmente será usável. Ele pode ser muito lento, muito grande, desajeitado em uso, ou todos os três. A questão administrativa, não é se deve construir um sistema-piloto e jogá-lo fora. Isso será feito. A única questão é se deve planejar antecipadamente a construção de algo que se vai jogar fora ou prometer entregar isso aos clientes. (PRESSMAN, 2005).

8 TRABALHOS CORRELATOS

A identificação de conceitos baseada em taxonomias para determinar tópicos de documentos é bastante explorada em diversos trabalhos, por exemplo, Lin (1995) adota essa estratégia para criar sumários de textos. Ele estende a contagem de palavras, como maneira de identificar os tópicos de um texto, para a contagem de conceitos, propondo um método para identificar automaticamente suas ideias centrais. Essa contagem é feita utilizando uma taxonomia para realizar generalizações, como, por exemplo, inferir que um texto que traga as palavras laptop e handheld pode tratar do tópico computadores portáteis. Para isso, é definido um peso para cada conceitoe esse peso representa a frequência de ocorrência dos itens lexicais. Tanto conceitos quanto subconceitos de itens lexicais podem ser usados para determinar esse peso. Além do peso, é definido o grau de generalização do conceito – variável G - calculada para cada conceito C da ontologia da seguinte maneira:

$$G_c = \frac{\text{MAX}_c(\text{ maior peso entre todos os subconceitos de C})}{\text{SUM}_c(\text{soma dos pesos de todos os subconceitos de C})}$$

Essa expressão indica que, quanto maior o valor de G, mais o superconceito C reflete um único subconceito, ou seja, o subconceito de C de maior peso, tem proporcionalmente um peso maior que os outros subconceitos. Opostamente, quanto menor o valor de G, maior será o equilíbrio entre os pesos dos subconceitos de C, e, portanto, C generaliza seus subconceitos. Nesta situação, se fosse escolhido um dos subconceitos de C como conceito principal, seriam perdidas informações, já que todos os subconceitos apresentam pesos similares e, portanto, são igualmente relevantes, devendo o superconceito ser escolhido como conceito principal. Por exemplo, suponha que, para certo texto, um conceito chamado “empresas” seja superconceito C dos seguintes subconceitos Ci (respectivos pesos entre parênteses):Toshiba(0), NEC(1), Compaq(1), Apple(7) e IBM(1). O peso final do conceito “empresas” e o seu valor G serão, respectivamente, 10 (0+1+1+7+1) e 0.70 (7/10). Neste exemplo, pode-se supor, com base no valor de G, que o subconceitoApple é o conceito principal do texto, pois ele é o conceito mais mencionado e que mais influiem G, cujo valor é 0.70. Segundo Lin (1995), o valor 0.68 (empiricamente determinado) serve como um limitante superior para G. Esse limite serve para indicar quais conceitos são considerados

relevantes no documento. Assim, se o valor de G estiver abaixo de 0.68, ele é considerado um conceito importante dentro do texto.

Com o objetivo de testar esse modelo, um experimento foi realizado envolvendo a sumarização de artigos da Business Week (1993-1994). A coleção de testes foi formada por 50 artigos sobre processamento de informação, com tamanho médio de 750 palavras. As medidas utilizadas para verificaro desempenho foram cobertura e precisão e a WordNet foi utilizada como ontologia. (MILLER, 1995). Para cada texto foi obtido um abstract (7-8 sentenças) feito por um profissional. Adicionalmente, foram construídos manualmente seus extratos ideais (com 8sentenças) pela extração e justaposição de sentenças que continham os conceitos principais mencionados nos abstracts. Os extratos ideais foram então comparados com aqueles gerados automaticamente, cujas sentenças foram pontuadas eselecionadas considerando três variações: 1) a pontuação de uma sentença corresponde à soma dos pesos dos conceitos-pais das palavras presentes na sentença; 2) o pesode uma sentença corresponde à soma dos pesos dos conceitos na própria sentença; 3) similar à variação 1, mas somente considerando um conceito (o mais relevante) por sentença. Os resultados obtidos (considerando extratos automáticos de 8 sentenças) apresentaram respectivamente, os seguintes valores de precisão (P) e cobertura (R): variação 1 (P=0.37, R=0.32), variação 2 (P=0.34, R=0.30), variação 3 (P=0.33, R=0.28).

Usando uma metodologia distinta de sumarização, Wu e Liu (2003) utilizaram artigos doThe New York Times e do Wall Street Journal como corpus para elaborar e construir a taxonomia utilizada na sumarização dos documentos, que é codificada como uma estrutura em árvore onde cada nó representa um conceito. Todos os artigos possuem extratos ideais e estes, por sua vez, têm parágrafos como unidade básica. Parágrafos foram escolhidos como unidades básicas para possibilitar a comparação entre os extratos gerados por este método e os extratos ideais do corpus que usam essa granularidade. Um processo de mapeamento verifica a correspondência entre as palavras do texto e os conceitos taxonômicos, atribuindo-lhes pesos. O peso de cada conceito em relação ao texto a ser sumarizado é calculado somando-se a frequência das palavras que aparecem no documento e que correspondam ao conceito. Aqui, cada palavra tem correspondência unívoca comcada conceito da taxonomia. Considerando a estrutura arbórea, quando o peso deum nó é incrementado, o incremento é propagado para seus ancestrais. Após rotular toda a árvore com os pesos, os nós de segundo nível da árvore (logo abaixo da raiz) com maiores pesos são considerados os tópicos principais do documento. Pelo fato de a taxonomia estar organizada como uma árvore com uma única raiz principal, e de se propagarem os pesos, a raiz da árvore (primeiro nível) receberia sempre o

maior peso. Além disso, representaria um único conceito muito genérico. Desta forma, escolher os nós do segundo nível garante que diferentes conceitos com diferentes pesos sejam considerados. Posteriormente, os parágrafos que tiverem maior proximidade com esses conceitos são selecionados.

A seleção de parágrafos é feita pontuando-os de acordo com a presença de palavras que correspondem aos conceitos principais identificados: para cada palavra relacionada a um conceito identificado anteriormente, o parágrafo recebe uma quantidade de pontos relativa ao peso do conceito considerado. Por exemplo, suponha a existência de um conceito chamado "cinema" cujo peso seja 20. Um parágrafo obterá 20 pontos para cada palavra que ele contenha e que seja relacionada a este conceito, por exemplo, as palavras filme e Spider-man. Essa abordagem foi testada para verificar o seu potencial. As medidas utilizadas foram precisão e cobertura. Elas foram calculadas verificando se as sentenças presentes no extrato automático correspondiam àquelas de um extrato ideal. Para o experimento foi utilizado um corpus de 51 artigos (com extratos ideais), num total de 882 parágrafos (com1 ou 2 sentenças), dos quais 133 faziam parte dos extratos ideais. Os resultados obtidos indicam que este método alcançou uma precisão que varia entre 0.24 (extratos com 10 parágrafos) e 0.70 (extratos com 1 parágrafo) e uma cobertura variando entre 0.94 (extratos com 10 parágrafos) e 0.70 (extratos com 1 parágrafo). Segundo Wu e Liu (2003), esses resultados mostram que a ontologia consegue captar informações relevantes para a tarefa de sumarização.

A pesquisa desenvolvida por Tiun (2001) explora a taxonomia do Yahoo para determinação do tópico principal de um documento. Um conjunto de palavras-chave é extraído de sentenças significativas do documento e, posteriormente, é mapeado no conjunto de seus conceitos ontológicos. As sentenças significativas são indicadas por um módulo de extração, que se baseia nas etiquetas HTML do documento (por exemplo, extraindo palavras presentes nos textos-âncora doslinks, palavras enfatizadas por itálico ou negrito ou marcadas como título do documento). As palavras-chave são, então, mapeadas em conceitos da taxonomia. Essa correspondência é feita comparando-se cada uma delas a um conjunto de itens lexicais que estão associados a cada conceito da taxonomia. Esse conjunto é construído juntamente com a taxonomia; na medida em que os conceitos taxonômicos são determinados, os itens lexicais que descrevem esses conceitos são associados. Inicialmente cada conceito tem associado a si um pequeno número de descritores que são extraídos diretamente do seu identificador. Por exemplo, um conceito ontológico do Yahoo denominado Artsand Humanities terá o item lexicalArts e o item lexicalHumanities como descritores associados.

O mapeamento entre palavras-chave e conceitos visadefinir o peso de cada conceito no documento. Esse peso é calculado pela soma da frequência das palavras-chave, no texto, coincidentes com os itens lexicais que descrevem o conceito e com a forma como o mapeamento é realizado. Palavras-chave mapeadas alternativamente por meio de um vocabulário estendido contribuirão com 50% da sua frequência para o peso de um conceito. O peso acumulado final de um conceito será o seu peso somado ao peso acumulado total de seu(s) conceito(s) filho(s), multiplicado pelo número de palavras-chave mapeadas sem utilização do vocabulário estendido. A pesquisa indica, para um conjunto de 202 documentos, uma precisão de 29.7% na determinação de seus tópicos principais. Explorando a identificação de conceitos para a classificação de páginas Web, Mladenic e Grobelnik (1999) também utilizam como base a ontologia do Yahoo para o inglês. Eles têm como hipótese que documentos de texto podem ser caracterizados por um conjunto de palavras-chave que permitem indicar seu conteúdo. Os documentos são representados como um vetor de características e incluem, além de unigramas, seqüências de até cinco palavras (5-gramas), ou pentagramas.

Inicialmente os conceitos da ontologia são descritos por meio de palavras extraídas das categorias definidas no Yahoo. Por exemplo, a categoria “Machine Learning” que, por sua vez, é uma subcategoria de “Science”, está hierarquicamente subordinada do seguinte modo: “Science: Computer Science: Artificial Intelligence: Machine Learning”. Assim, o conceito Machine Learning tem associado a si, como descritores, as palavras-chave: Science, Computer Science, Artificial Intelligence e Machine Learning. Técnicas de aprendizado de máquina (Naive-Bayes) são utilizadas usando como dados de treinamento documentos previamente classificados segundo essa hierarquia. O problema de classificação é dividido em subproblemas: para cada conceito da ontologia existe um classificador associado, que visa reconhecer documentos relacionados àquele conceito. O resultado é um conjunto de classificadores independentes que, em conjunto, são usados para determinar o tópico principal de um documento pela associação de suas palavras-chave aos conceitos da ontologia. Isso é feito computando-se sua similaridade ou a probabilidade de a palavra-chave pertencer a uma determinada classe conceitual.

Essa abordagem foi testada por meio de um experimento cujo objetivo foi verificar se um novo documento apresentado era corretamente classificado. Como medida de desempenho, foi realizada uma verificação simples, checando para quais classes conceituais o documento era indicado com maior probabilidade. O conjunto de testes foi formado por 1.100 documentos, os resultados experimentais mostraram que, para cerca de 50% dos exemplos de

testes, o conceito correto estava entre os três de maiores probabilidades apontados pelos classificadores.

De modo geral, os trabalhos apresentados demonstram o potencial de utilização de conhecimento ontológico para a detecção de tópicos em texto. A motivação comum a todos eles é a de que a análise de palavras isoladas em um texto, sem nenhuma consideração sobre o relacionamento semântico entreelas pode ser um fator limitante no processamento de documentos. Desta forma, todos eles adotam como elemento fundamental uma taxonomia. Outras similaridades podem ser apontadas como, por exemplo, processos de mapeamento que associam conceitos aos documentos verificando a correspondência entre as palavras do texto e a taxonomia. Processos de propagação de peso entre os conceitos com o intuito de realizar generalizações também são pontos em comum entre os trabalhos citados. Todos esses aspectos foram observados e considerados para a elaboração da proposta apresentada nesta monografia.

9 METODOLOGIA

O propósito das pesquisas exploratórias é proporcionar ao investigador maior familiaridade com o problema, objetivando torná-lo mais explícito ou construir hipóteses.

Dessa forma, este projeto é uma pesquisa exploratória, pois visa investigar métodos de construção de um software para detecção automática de perfis de usuários com base em documentos textuais, utilizando teorias associadas à área de Processamento de Língua Natural e Inteligência Artificial.

Para construção de um sistema dessa natureza pode-se observar uma série de etapas que devem ser seguidas até a validação do protótipo.

O processo de desenvolvimento da ferramenta proposta deve iniciar com a definição da estrutura do projeto e a definição detalhada de metas e cronogramas. Um ponto importante no estabelecimento do projeto do sistema é a decisão sobre que informação relevante deve ser mostrada ao usuário final, neste caso, os perfis dos usuarios que estão sendo analisados.

Fase 1:

1.1. Seleção de uma ontologia.
1.2. Enriquecimento da ontologia selecionada.
1.3. Desenvolvimento da ferramenta de detecção de tópicos.
1.4. Arquitetura e funcionamento do EXTRATOP.
1.5. Avaliação da ferramenta EXTRATOP.
1.6. Considerações sobre a avaliação do EXTRATOP.

Fase 2:

2.2. Revisão da ontologia
2.3. Prototipação da ferramenta
2.4. Arquitetura
2.5. Seleção de dados
2.6. Limpeza de dados
2.7. Avaliação do processo
2.8. Execução e classificação
2.9 Resuultado da avaliação da ferramenta AboutYou

Onde a Fase 1, basicamente, é constituída na construção e enriquecimento da ontologia num âmbito geral, em que o foco é a detecção de assuntos de um texto e não a detecção de perfis.

A Fase 2 é onde há a construção de uma ferramenta de detecção de perfil (AboutYou), utilizando como base os recursos da Fase 1.

9.1 FASE 1

9.1.1 Seleção de uma ontologia

A primeira etapa do desenvolvimento deste trabalho consistiu em um estudo das tecnologias disponíveis para a especificação e representação da ontologia e utilização do algoritmo de seleção de tópicos, que serão usados para definir o perfil de um usuário.

O sistema faz a identificação de tópicos pela contagem de conceitos, usando a ontologia do Yahoo Enriquecida adaptada de (PEDREIRA-SILVA, 2006).

A razão da escolha da ontologia do Yahoo para esta pesquisa se deu por várias razões: (1) dentro da história da Internet o Yahoo foi durante muito tempo uma das principais ferramentas usadas por internautas para encontrar informações hierarquicamente organizadas em categorias na Internet; (2) essa ontologia é uma das maiores já compiladas por humanos: seu conteúdo foi organizado por editores que visitavam, analisavam e incluíam sites, organizando-os em categorias de acordo com o assunto; (3) seu conteúdo está disponível também em língua portuguesa; (4) a ontologia enriquecida permite sua utilização quase direta por sistemas de processamento de língua natural.

As principais categorias, ou conceitos, do Yahoo incluem: Artes e Cultura, Esportes, Educação, Ciência, Regional, Business to Business, Fontes de Referência, Saúde, Compras e Serviços, Lazer, Informática, Internet, Notícias, Finanças, Governo e Sociedade. Cada conceito é descrito por um conjunto de palavras-chave que o caracteriza. As palavras-chave, por sua vez, delineiam um caminho que indica a posição do conceito na hierarquia. Em outras palavras, um subconceito é descrito adicionando uma palavra-chave ao conjunto de palavras-chave que caracterizam seu superconceito. Considerando essa sucessão de atribuição de conceitos a cada nó da hierarquia, todos os nós da hierarquia herdarão de forma crescente os conceitos de seus sucessores. Um exemplo de caminho com cinco conceitos inter-relacionados é indicado por “>>”, como segue. O superconceito, neste caso, é Artes e Cultura e o subconceito mais elementar ou folha é Bibi Ferreira:

Artes e Cultura>>Artes Cênicas>>Artistas>>Atores e Atrizes>>Bibi Ferreira

São os conceitos como esses que são utilizados pela ferramenta implementada como possíveis tópicos (assuntos) de um documento que caracterize um usuário nesta pesquisa. Porém, conforme descreve Pedreira-Silva (2006) a incorporação da ontologia do Yahoo por um sistema computacional não se dá de forma direta, justificando o processo de enriquecimento realizado pelo autor. Tal processo, que será descrito a seguir, culminou na geração de uma base de dados ontológica que foi incorporada ao sistema proposto nesta iniciação científica.

A descrição de conceitos na ontologia do Yahoo originalmente não segue um modelo específico. Os itens lexicais utilizados para descrever cada conceito apresentam variações com relação às suas formas. São encontrados, por exemplo, itens que descrevem os conceitos e que são formados por palavras isoladas (como artesanato, dança, design) ou composições de palavras (por exemplo,Cinema e Filmes ou Centros Culturais). Variações de número e gênero também podem ser percebidas, como nos conceitos “artistas”, “Atores e Atrizes” e “dança”. Tais características guiaram, portanto, o processo de enriquecimento realizado e descrito por (PEDREIRA-SILVA, 2006) que se justifica pelo fato de a ontologia do Yahoo ser considerada pobre (CHEN, 1994; CHEN et al., 1997; FURNAS et al., 1987; TIUN et al., 2001) para dar conta do vocabulário de textos livres (isto é, textos de autoria). Esse problema ocorre porque o processo de identificação de tópicos de textos envolve o mapeamento das palavras da língua natural para conceitos ontológicos, assim, diferentes palavras se referem muitas vezes ao mesmo conceito ou vice-versa.

Claramente fazer esse mapeamento entre conceitos e palavras é extremamente difícil já que a língua natural permite uma série de variações devido aos sinônimos (palavras diferentes com o mesmo significado), à polissemia (a mesma palavra com diferentes significados), às variações léxicas (uso de radicais, conjugações verbais, variações de gênero e número) e aos chamados quase-sinônimos (palavras correlatas, como "bomba" e "explosão"). O conceito “Artes e Cultura”,por exemplo, poderia ter como descritores diversas palavras: artes, cultura, “artes e cultura”, música, artesanato, etc. Define-se como enriquecimento, no contexto desta pesquisa, o processo de descrever um conceito da ontologia do Yahoo por meio de palavras da língua natural em foco). A metodologia aplicada por Pedreira-Silva (2006) envolveu a coleta manual de um vocabulário externo, utilizando como descritores de

conceitos palavras que tenham algum tipo de relação semântica com os conceitos ontológicos. Estas relações podem ser de diversos tipos, por exemplo, sinonímia, hiponímia e hiperonímia.

Este enriquecimento permitiu aumentar o poder de generalização da ontologia (por exemplo, dizer que Bogotá e Medellin remetem ao superconceito derivado do país comum – Colômbia). Por questões de simplicidade, o enriquecimento da ontologia do Yahoo para português foi realizado restringindo-se a descrição dos conceitos às palavras encontradas no thesaurus Diadorim e na Wikipédia. Itens lexicais coletados corresponderam às palavras de classe aberta (substantivos, verbos e adjetivos), convertidos manualmente para a forma canônica. Além da forma canônica, para substantivos e adjetivos foram incluídas as variações léxicas mais comuns (forma feminina e plural). Desta forma, um único conceito da ontologia passou ser descrito por itens lexicais diferentes. Por exemplo, o conceito denominado Atleta foi descrito por “atleta”, “atletas”, “desportista”, “esportista” etc.

Em um primeiro passo de enriquecimento, foram utilizadas como descritores as palavras que identificavam os conceitos na ontologia e, posteriormente, foi considerada a relação semântica de sinonímia, para completar o enriquecimento da ontologia original. Em uma segunda etapa de enriquecimento, utilizou-se os próprios documentos da Internet como fontes externas de conhecimento, para extrair deles os subsídios para enriquecer a ontologia. Mais particularmente, a fonte de conhecimento usada foi a Wikipédia, uma enciclopédia livre em construção por milhares de colaboradores de todo o mundo.

O enriquecimento da ontologia do Yahoo foi feito da seguinte forma: para cada conceito foram coletados manualmente, diretamente da Wikipédia, documentos cuja temática tinha relação com a ideia expressa pelo conceito. Uma vez recuperados, a seleção das informações relevantes para o enriquecimento foi feita manualmente por um engenheiro do conhecimento. Apesar do caráter subjetivo da intervenção humana na descrição de conceitos alguns trabalhos, por exemplo, (LOH, 2001), sugerem que essa intervenção pode facilitar o processo de descrição de conceitos e melhorar os resultados finais. Após a leitura dos documentos recuperados, o engenheiro de conhecimento escolheu as palavras que melhor descreviam os conceitos já existentes na ontologia do Yahoo, relacionando-os à ontologia. Assim, procurou-se garantir um conjunto mínimo de 26.300 descritores acrescentados a essa ontologia (PEDREIRA-SILVA, 2006). Cabe destacar que esses descritores podiam ser compostos por mais de uma palavra, como por exemplo, “Bibi Ferreira”.

O enriquecimento da ontologia do Yahoo para o português, por meio de duas fontes de natureza distinta – a Wikipédia e o Diadorim – teve um aspecto positivo no processo de generalização de conceitos, o que justifica a escolha deste recurso para esta pesquisa, já que o

objetivo é justamente implementar uma ferramenta computacional que seja capaz de generalizar e apontar automaticamente os tópicos principais de um documento. Ambos os recursos são complementares, já que a Wikipédia é enciclopédica, enquanto o Diadorim agrega somente informações paratáticas de sinonímia e antonímia. Estas, claramente, não permitem generalizações. Por exemplo, a utilização pura e simples de um thesaurus, sem o tipo de léxico agregado pelo engenheiro de conhecimento, usando uma fonte como a Wikipédia, tornaria impossível generalizar e determinar, por exemplo, o conceito restaurante a partir de palavras como "garçom", "cliente", "comida" e "menu". Neste exemplo, as palavras "garçom", "cliente", "comida" e "menu" remetem claramente a um local que é expresso por meio do conceito restaurante. Esse tipo de informação poderia ser coletado de um texto na Wikipédia que tratasse, por exemplo, da origem e da definição da palavra restaurante. Já othesaurus poderia acrescentar informações mais limitadas, como por exemplo, indicar que a palavra “menu” é sinônimo de “cardápio”.

Atualmente, a ontologia enriquecida por esse processo possui aproximadamente 5.500 conceitos e cerca de 26.300 descritores associados aos conceitos, utilizando a técnica de enriquecimento descrita (PEDREIRA-SILVA, 2006). Cabe destacar que o enriquecimento foi feito para 2.500 conceitos originais da ontologia do Yahoo (aproximadamente metade da coleção).

9.1.2 Enriquecimento da ontologia selecionada

Nesta pesquisa, adicionalmente foram acrescentados de forma manual um pequeno conjunto de novos descritores para a ontologia de forma a atualizá-la e torná-la mais rica.

Entretanto, como o foco desta pesquisa está na verificação do potencial uso da ontologia em tarefas de detecção automática de perfis, não houve preocupação em completar essa coleção devido ao esforço de se realizar esse enriquecimento manualmente. Esse repositório será agregado à uma ferramenta denominada AboutYou cujo processo de desenvolvimento será explicitado na seção seguinte e que foi uma das atividades desta pesquisa.

9.1.3 Desenvolvimento da ferramenta de detecção de tópicos.

Após terem sido recolhidos todos os conhecimentos e informações relacionadas para a execução do trabalho referente à ontologia, foi iniciada a fase de desenvolvimento do protótipo EXTRATOP, cuja função é detectar tópicos que serão usados para definir um possível perfil de um usuário. Este desenvolvimento consistiu principalmente na implementação de um banco de dados de contendo as informações de natureza ontológica, do desenvolvimento do algoritmo de pontuação de tópicos e da implementação de uma interface Web.

A criação do protótipo correspondeu ao estágio de desenvolvimento inicial, onde o valor do sistema foi avaliado, considerando testes iniciais executados pelo próprio autor deste trabalho. Estes testes iniciais foram úteis para verificar se as funcionalidades estavam aceitáveis para prosseguir com a pesquisa ou se havia necessitaria de alguns ajustes ou melhorias.

Os resultados obtidos dessa avaliação inicial do protótipo serviram como base para melhorias da ferramenta, permitindo o aprimoramento dos seus algoritmos.

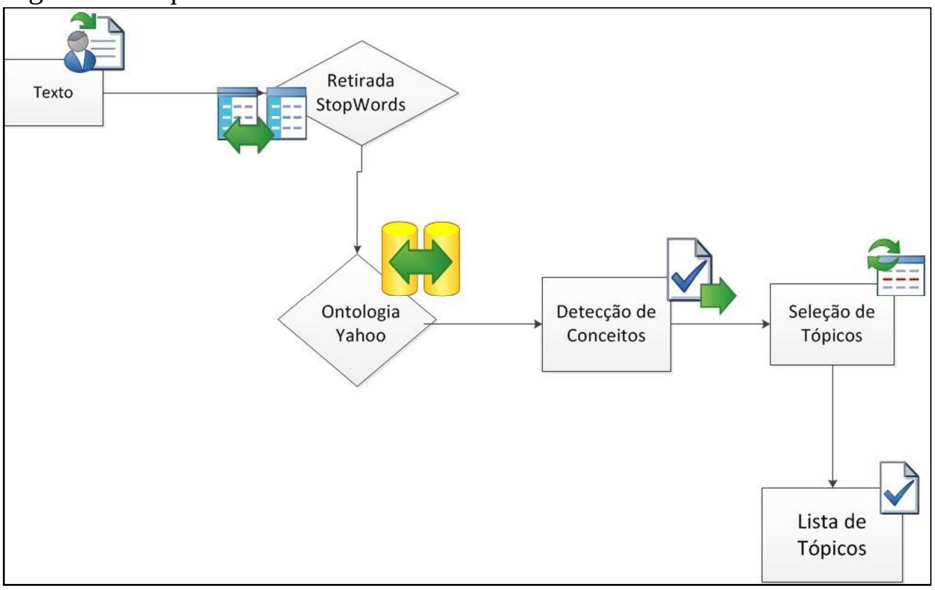
O software foi desenvolvido utilizando a linguagem de programação PHP, JavaScript e banco de dados MySQL. Foi utilizado o software Sublime Text como ambiente de desenvolvimento. A seção a seguir trata da arquitetura e do funcionamento do EXTRATOP.

9.1.4 Arquitetura e funcionamento do extratop

Na primeira fase da investigação um protótipo da ferramenta foi gerado e, posteriormente, alguns testes (descritos na seção seguinte) foram realizados, visando verificar a sua eficiência para a tarefa de detecção de tópicos.

A arquitetura da ferramenta engloba a parte de interface com o usuário e a definição do algoritmo de detecção de tópicos a ser utilizado. A arquitetura do EXTRATOP é exibida na Figura 4.

Figura 4 – Arquitetura do EXTRATOP



Fonte: Elaborada pelo autor

De acordo com esta arquitetura, um texto-fonte é dado como entrada para o sistema e os seguintes passos são realizados para detecção dos tópicos:

1. Inicialmente o texto-fonte é pré-processado, sendo retirados os sinais de pontuação tradicionais (por exemplo, ponto final, ponto de exclamação e ponto de interrogação);

2. Conceitos subjacentes ao texto são detectados com base na ontologia do Yahoo conforme será descrito na próxima seção.

3. As stopwords do texto são removidas e, então, os tópicos são ranqueados e selecionadas de acordo sua proximidade com os conceitos da ontologia.

4. Finalmente, são exibidos como resposta final os tópicos que indicam o conteúdo principal do documento.

O primeiro passo do método de detecção de tópicos proposto é determinar aqueles que são mais importantes no documento, estimando sua relevância. Isto é feito verificando se as palavras presentes no texto correspondem àquelas que descrevem os conceitos ontológicos. Esse procedimento de verificação de correspondência é o que define-se como mapeamento. Sempre que essa correspondência ocorrer, assume-se que aquele conceito é subjacente ao texto e representa, portanto, um de seus tópicos. O processo de mapeamento ocorre após o pré-processamento do texto que inclui a retirada das chamadas stopwords que correspondem às palavras que em uma busca podem ser consideradas irrelevantes (os, as, de, do,...).

Especificamente para esta pesquisa foi elaborada uma lista com cerca de 255stopwords (a lista completa encontra-se nos anexos deste relatório).

Como fator de discriminação da importância dos conceitos, inicialmente é calculado o peso de todos eles. O cálculo do peso de um conceito é feito tendo por base a frequênciadas palavras no documento que são mapeadas no conceito, ou seja, que correspondem aos descritores dos conceitos. A frequência é identificada pela contagem absoluta das ocorrências da palavra no documento. O cálculo do peso de um conceito, com base na frequência das palavras, parte do princípio de que a repetição de palavras em um texto é feita com o intuito de enfatizar algum assunto e pode ser um indicador de significância das palavras (SALTON;MACGILL, 1983). No EXTRATOP, sempre que uma palavra do texto corresponder a um descritor de um conceito, o peso deste conceito é incrementado em 1 unidade; esse processo é cumulativo, ou seja, toda vez que a mesma palavra aparece no texto adiciona-se 1 unidade. Considerando a estrutura hierárquica da ontologia e os relacionamentos entre os conceitos, a detecção de um conceito subjacente ao documento implica indiretamente a presença de seu conceito-pai também no documento (TIUN et al., 2001). Por exemplo, se considerarmos uma relação ontológica entre os conceitos Futebol e Esporte, em que Esporte é pai de Futebol, a presença do conceito Futebol em um texto indica que, em um nível mais genérico, o conceito Esporte também está presente no texto.

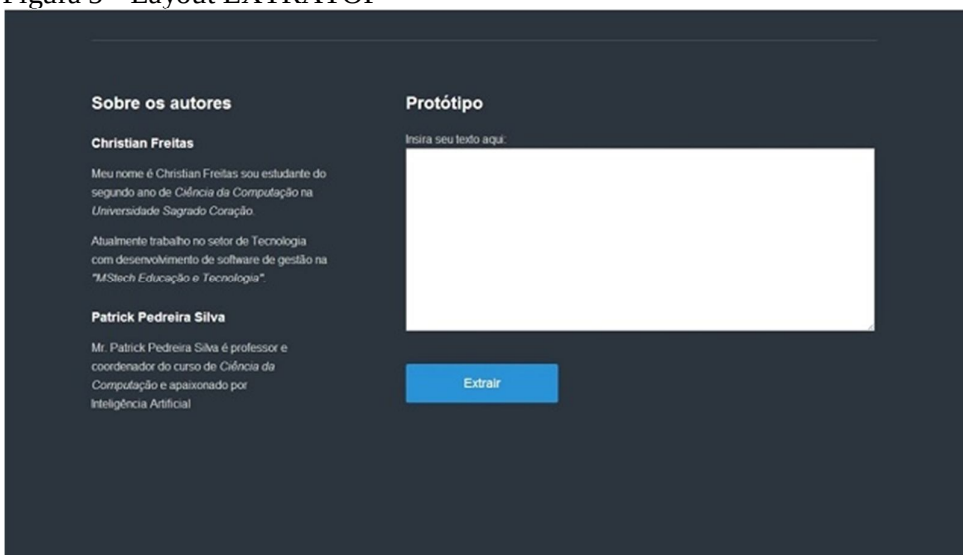
Com base no exposto, para tentar modelar esse processo de generalização em que um conceito é detectado a partir de seus conceitos-filhos, na sequência desta pesquisa será implementado um processo de pontuação de tópicos que considera que sempre que o peso de um conceito é incrementado devido ao mapeamento de uma palavra, o peso de seu pai também é incrementado. Cabe destacar que, no contexto considerado nesta investigação, esse processode generalização é considerado somente de conceito-filho para seu conceito-pai (contexto imediato). A decisão de considerarsomente o contexto imediato foi tomada em virtude da própria estrutura da ontologia do Yahoo. Conforme acusa Fenselet al. (2002), a ontologia do Yahoo provê uma noção básica de generalização e especialização com a maior parte da relações do tipo "é-um". Por exemplo, considerando uma relação entre conceitos indicada por >>, onde o conceito da esquerda é o pai do conceito da direita, em Esporte>>Artes Marciais>>Capoeira, temos que Capoeira é uma Arte Marcial. No esquema geral de organização da ontologia do Yahoo, tipicamente um conceito-filho de um conceito mais específico também mantém a relação do tipo "é-um" com os conceitos mais genéricos (para o exemplo anterior, podemos considerar que Capoeira é também um Esporte). Entretanto, nem todas as relações desta ontologia seguem

estritamente esse tipo derelação "é-um", fazendo com que o processo de generalização que proposto não seja aplicável sem considerar um limite. Por exemplo, considerando um conjunto de conceitos relacionados na ontologia, indicados por Vestuário >> Vestuário Feminino >> Acessórios Femininos >> Maquiagem; podemos considerar que Maquiagem é um tipo de Acessório Feminino, mas já não poderíamos dizer que é um tipo de Vestuário. Adicionalmente a esse problema, devido ao grande número de conceitos, a propagação para todos os níveis da ontologia tornaria o processo computacionalmente ineficiente.

Considerando que o mapeamento é um processo recorrente sobre a estrutura hierárquica da ontologia, por esse procedimento de propagação de pesos os conceitos mais próximos da raiz obteriam os maiores pesos e claramente os conceitos terminais seriam prejudicados. Com o intuito de evitar essa situação, será feita a propagação de pesos entre conceitos pais e filhos de modo similar à usada por Tiun et al. (2001). Assim, sempre que um peso é propagado do conceito-filho para o conceito-pai seu valor é reduzido. No EXTRATOP, a propagação de pontos para o conceito-pai corresponderá a 50% do peso obtido pelo conceito-filho. Esse valor de redução foi empiricamente escolhido e incluído no algoritmo da ferramenta.

As Figuras 5 e 6 mostram a interface do protótipo desenvolvido.

Figura 5 – Layout EXTRATOP



Fonte: Elaborada pelo autor.

Figura 6 – Layout de apresentação.

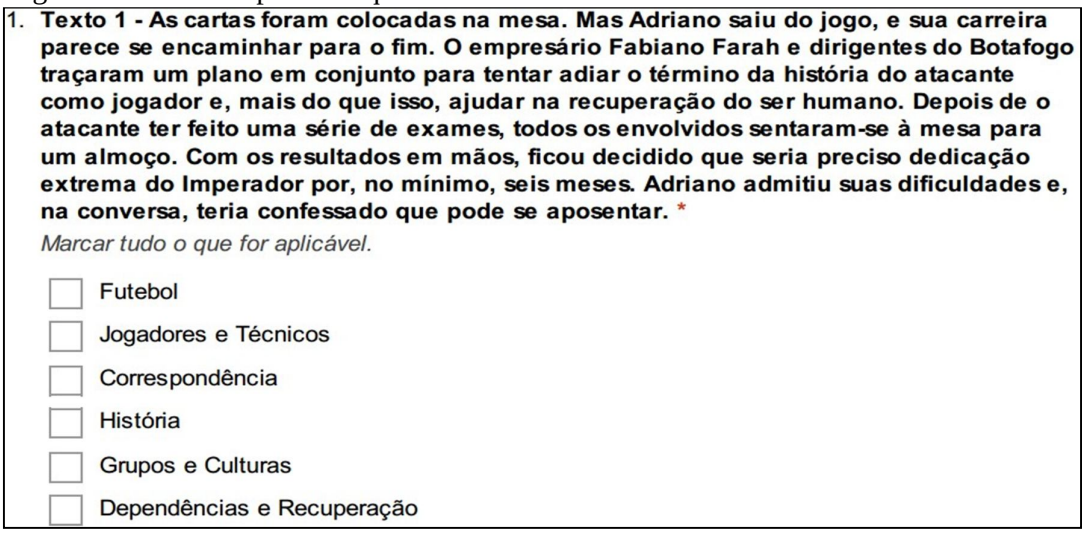
Fonte: Elaborada pelo autor.

9.1.5 Avaliação da ferramenta EXTRATOP

Como forma de testar o desempenho do protótipo na tarefa de detecção de tópicos relevantes foi elaborado um experimento que envolveu a participação de 40 voluntários. O experimento consistiu em apresentar, para cada voluntário, um conjunto de 5 textos curtos, cada um sobre um tema específico (Futebol, Musica, Educação, Economia e Política). Associada a cada texto havia uma lista com 6 ou 5 tópicos (definidos automaticamente pela ferramenta como relevantes) que deveriam ser analisados e assinalados caso os voluntários entendessem que os mesmos representassem tópicos relacionados ao textos.

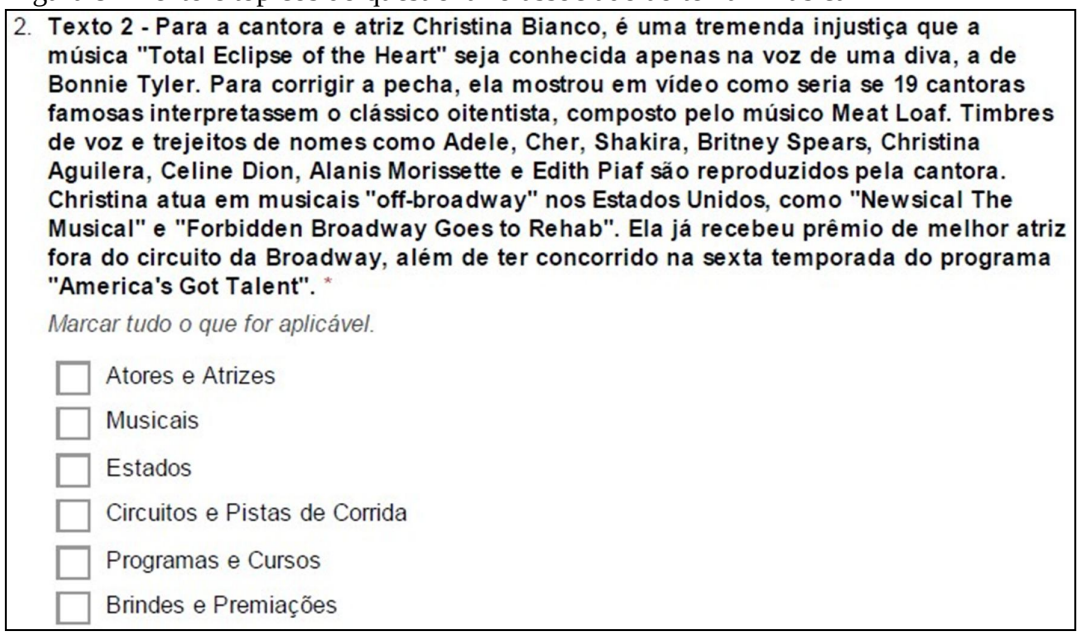
As figuras 7 e 8 mostram os textos utilizados no experimento e os respectivos tópicos associados.

Figura 7 – Texto e tópicos do questionário associado ao tema “Futebol”.



Fonte: Elaborada pelo autor.

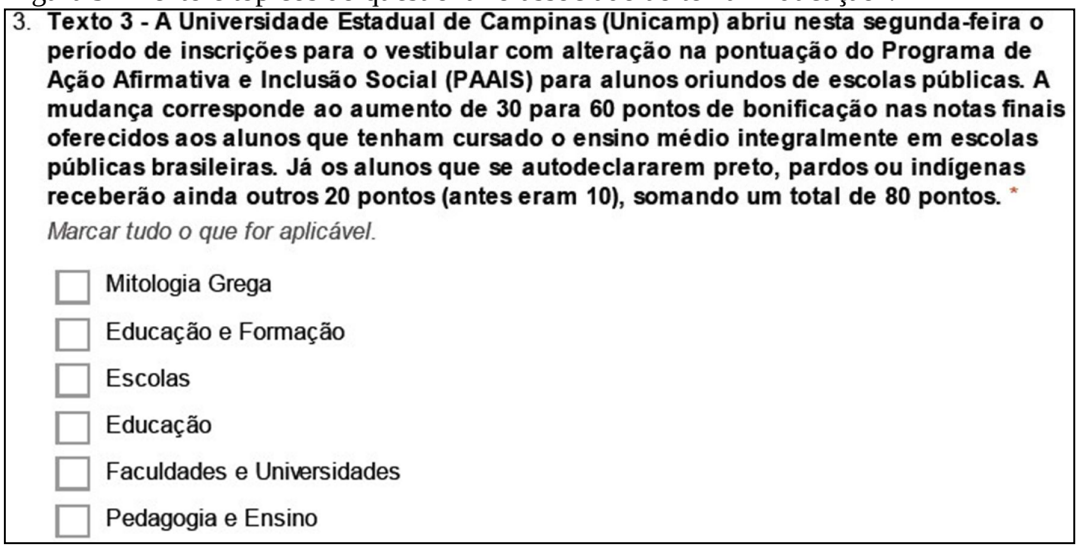
Figura 8 – Texto e tópicos do questionário associado ao tema “Música”



Fonte: Elaborada pelo autor.

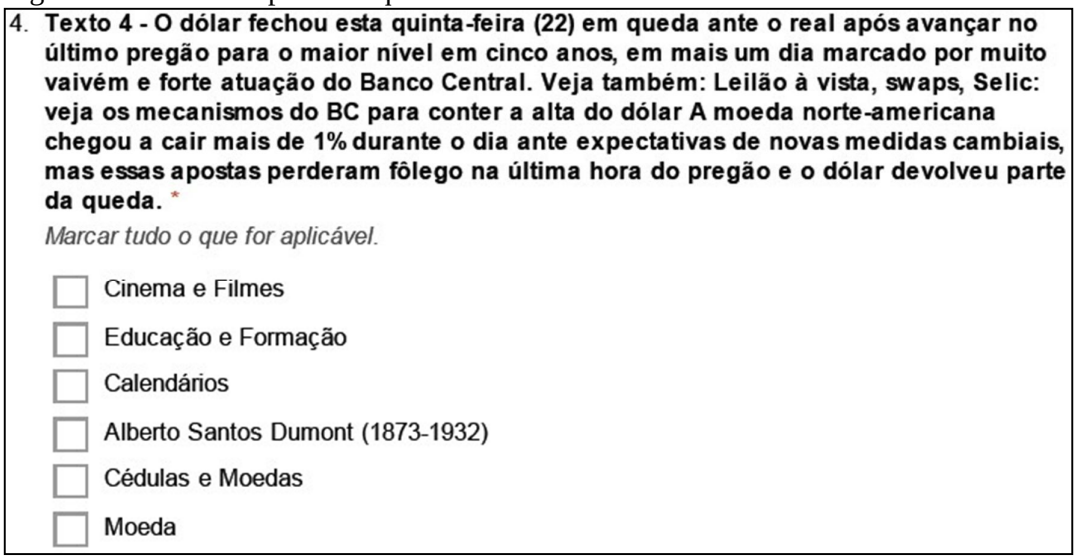
As figuras 9, 10 e 11 mostram os textos utilizados no experimento e os respectivos tópicos associados.

Figura 9 – Texto e tópicos do questionário associado ao tema “Educação".



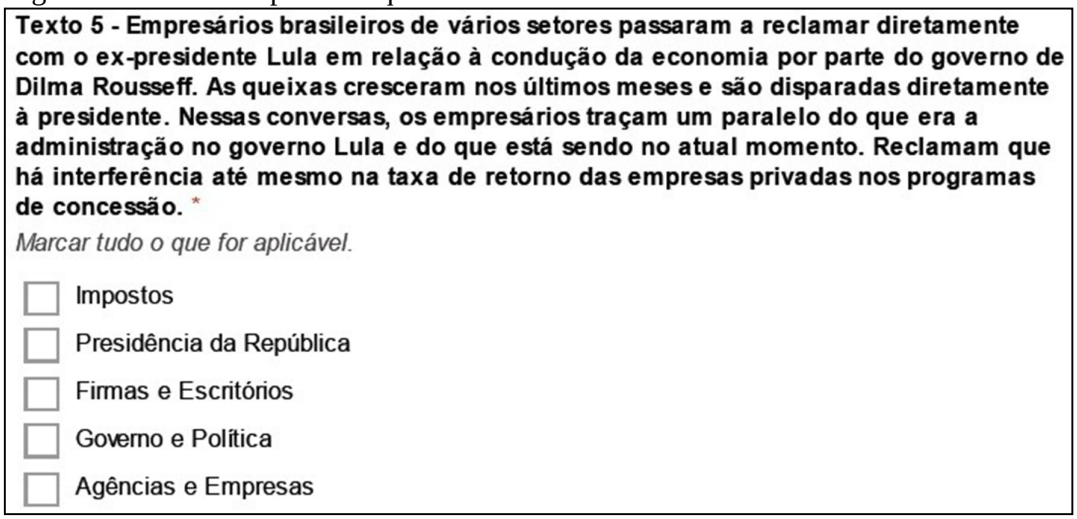
Fonte: Elaborada pelo autor.

Figura 10 – Texto e tópicos do questionário associado ao tema “Economia”.



Fonte: Elaborada pelo autor.

Figura 11 – Texto e tópicos do questionário associado ao tema “Política”.



Fonte: Elaborada pelo autor.

Para avaliar as respostas do protótipo foi verificado, para cada texto, qual o nível de concordância das pessoas com os tópicos apresentados. Apesar de este ser um experimento simples ele é útil para mostrar, em um primeiro momento, o potencial da abordagem implementada. Para esta análise foi definida como medida de precisão do sistema, a relação entre os 6 tópicos identificados pelo EXTRATOP para cada texto e aqueles que efetivamente foram referenciados pelos voluntários. A fórmula seguinte mostra como essa relação foi calculada:

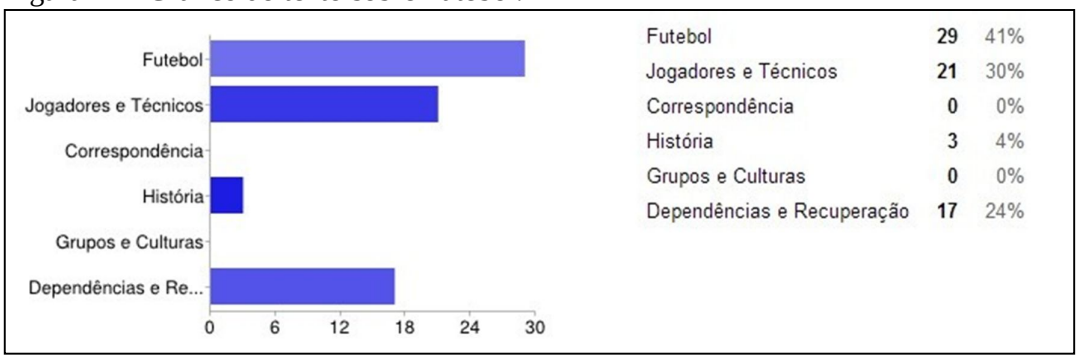
$$Precisão = \frac{número\ de\ tópicos\ considerados\ relevantes\ pelos\ usuários}{número\ de\ tópicos\ considerados\ relevantes\ pelo\ protótipo}$$

Neste sentido, quanto mais alta a precisão (valor mais próximo de 1) maior é a chance de o sistema estar indicando bons tópicos para o texto.

A figura 12 seguinte mostra os dados coletados referentes ao Texto 1 (Figura 7), relacionado ao tema Futebol. Na figura estão registradasquantas indicações cada um dos tópicos apresentados recebeu. Considerando,portanto, o primeiro texto pode-se notar que 4 dos 6 tópicos foram apontados também pelos usuários como sendo relevantes. Assim, a precisão do sistema ficou em torno de 67%. Uma observação interessante é que “Futebol” e

“Jogadorese Técnicos” foram os tópicos mais indicado pelos voluntários (41% e 30%, respectivamente) e estes são os tópicos mais relevantes apontados também pelo EXTRATOP.

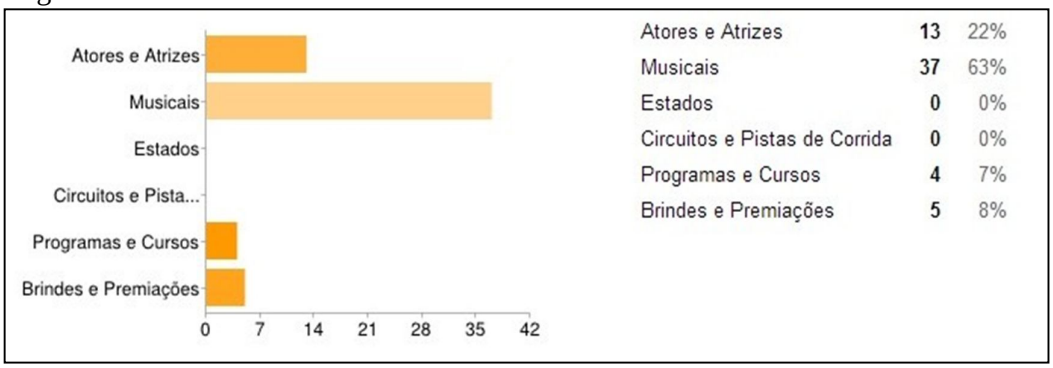
Figura 12 – Gráfico do texto sobre Futebol.



Fonte: Elaborada pelo autor.

A Figura 13 seguinte traz os dados referentes ao Texto 2 (Figura 8) sobre Música. De acordocom figura nota-se que 4 dos 6 tópicos foram apontados também pelos usuários como sendo relevantes. Desta forma, do mesmo modo que ocorreu com o Texto 1, a precisão do sistema foi de 67%. Da mesma forma que ocorreu com o outro texto, houve uma correspondênciaentre os dois tópicos mais indicados pelos voluntários e os tópicos mais relevantes apontados pelo EXTRATOP, no caso “Atores e Atrizes” e “Musicais”.

Figura 13 – Gráfico do texto sobre Música.

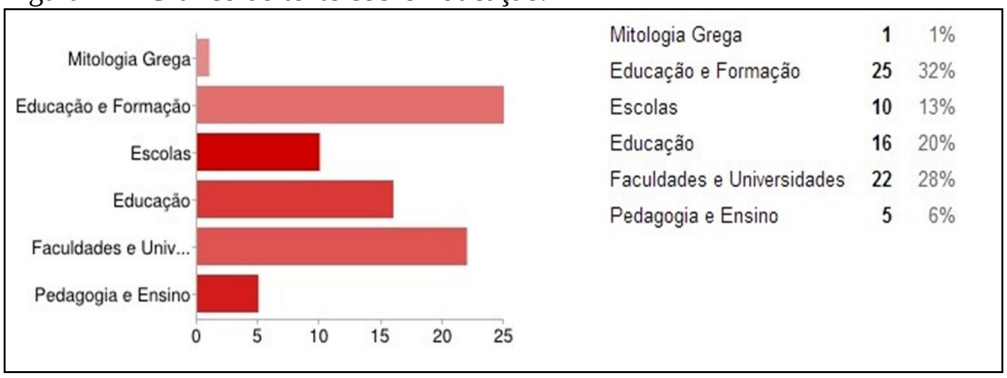


Fonte: Elaborada pelo autor.

A análise do resultado (Figura 14) referente ao Texto 3 (Figura 9), relacionado ao tema Educação, mostra que todos os tópicos foram assinalados pelos voluntários. Desse modo, aprecisão específica para esse texto ficou em 100%. Diferentementedo que aconteceu com outros textos, os resultados ficaram concentrados em mais de duas alternativas,

especificamente em quatro alternativas que somadas deram 93% das indicações (“Educação e Formação” com 32%, seguida por “Faculdade e Universidade” com 28%, “Educação” com 20% e “Escolas” com 13%). Observa-seque as alternativas mais marcadas possuem um assunto núcleo entre elas (Educação), que é também o tópico mais relevante apontado pelo EXTRATOP.

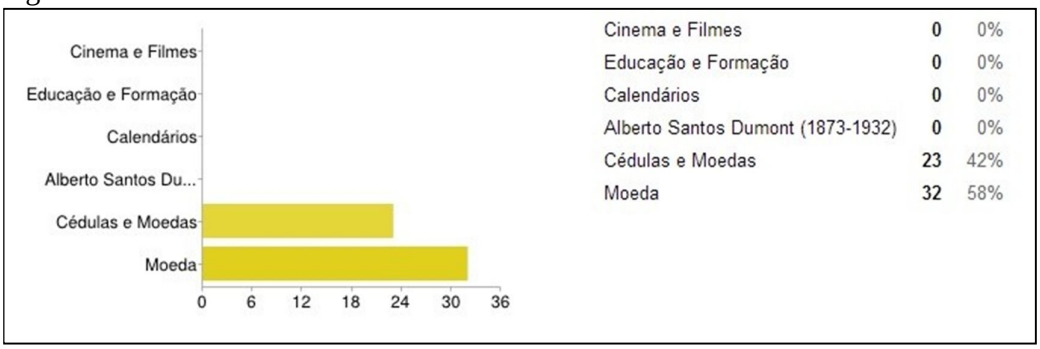
Figura 14 – Gráfico do texto sobre Educação.



Fonte: Elaborada pelo autor.

A Figura 15, referente ao Texto 4 (Figura 10) sobre Economia, mostra que os resultados tiveram uma particularidade, concentrando-se em apenas duas alternativas “Cédula e Moedas” e “Moedas”, respectivamente com 42% e 58% das indicações. Deste modo, a precisão do sistema ficou em 33%. Apesar da baixa precisão verificada, esse dois tópicos são justamente os dois principais apontados pela ferramenta EXTRATOP.

Figura 15 – Gráfico do texto sobre Economia.



Fonte: Elaborada pelo autor.

O último texto avaliado refere-se à Politica (Figura 16). A exemplo do que ocorreu com o Texto 3, a precisão ficou em 100%, uma vez que todos os tópicos foram apontados nas respostas dos voluntários. Nesse caso, os dois itens mais citados, segundo os entrevistados,

foram “Governo e Política” e “Presidencia e República”, respectativamente com 40% e 30%, coincidindo novamente com os dois tópicos mais relevantes apontados pelo EXTRATOP.

Figura 16 – Gráfico do texto sobre Política.

Fonte: Elaborada pelo autor.

9.1.6 Considerações sobre a avaliação do EXTRATOP

Numa análise geral de todos os resultados, verifica-se o protótipo obteve uma precisão média de 73,4%. Esse resultado é considerado satisfatório pois, de um modo geral, a ferramenta aponta para tópicos que estão, de fato, relacionados aos textos; o que pode indicar potencialidades na abordagem sugerida nesta investigação. Entretanto, muitos ajustes com relação ao conteúdo da ontologia devem ser feitos já que, pelo modo que a mesma foi estruturada, quando se considera um conjunto maior (acima de 4 ou 5 tópicos), os tópicos associados pela ferramenta aparentemente destoam do conteúdo dos textos como, por exemplo, no experimento realizado um dos tópicos associados ao texto sobre Economia foi “Cinema e Filmes”.

Isso ocorre devido ao fato de a língua portuguesa possuir várias palavras com mais de um sentido, gerando ambiguidade e, por lidar apenas de modo superficial com a linguagem, a ferramentanão faz nenhum tipo de desambiguação. Por outro lado, observa-se também que existem palavras presentes na ontologia que, por serem bem específicas de certos temas, ajudam a ferramenta a definir claramente os tópicos contidos nos textos. Isso gera uma diferençade resultados considerando-se as diferentes temáticas dos textos. Porexemplo, no experimento realizado os textos referentes à “Educação” e “Política” foram os que apresentaram a maior precisão. Isso é ocorre por dois motivos possíveis: primeiro, existem muitas palavras específicas que se referenciam diretamente e unicamente aos temas abordados ou então, os temas citados tem mais palavrasassociadas na ontologia, até mesmo porque não houve nenhum tipo de preocupação em balancear o número de termos associados a cada conceito.

9.2 FASE 2

Tomando como base a ferramenta EXTRATOP, algumas atividades foram desenvolvidas a fim de permitir a implementação da ferramenta ABOUTYOU. As atividades estão descritas nas seções seguintes.

9.2.1 Revisão da ontologia

Adicionalmente foi acrescentado de forma manual um pequeno conjunto de novos descritores (palavras) na ontologia de forma a atualizá-la e torná-la mais rica.

Entretanto, como o foco desta pesquisa está na verificação do potencial uso da ontologia em tarefas de detecção automática de perfis, não houve preocupação em completar essa coleção devido ao esforço de se realizar esse enriquecimento manualmente. Esse repositório será agregado à uma ferramenta denominada AboutYou cujo processo de desenvolvimento será explicitado na seção seguinte e que foi uma das atividades desta pesquisa.

9.2.2 Prototipação da ferramenta

Após terem sido recolhidos todos os conhecimentos e informações relacionadas para a execução do trabalho referente à ontologia emineração de dados foi iniciada a fase de desenvolvimento do protótipo AboutYou, cuja função é definir um perfil de um usuário através de sua rede social, nesse caso optou-se pelo Facebook. Este desenvolvimento consistiu principalmente na implementação de um banco de dados contendo as informações de natureza ontológica, do desenvolvimento do algoritmo de pontuação de tópicos de definição de perfis e da plataforma de extração de dados.

A criação do protótipo correspondeu ao estágio de desenvolvimento inicial, onde o valor do sistema foi avaliado, considerando testes iniciais executados pelo próprio autor deste trabalho. Estes testes iniciais foram úteis para verificar se as funcionalidades estavam aceitáveis para prosseguir com a pesquisa ou se havia necessitaria de alguns ajustes ou melhorias.

Os resultados obtidos dessa avaliação inicial do protótipo serviram como base para melhorias da ferramenta, permitindo o aprimoramento dos seus algoritmos.

O software foi desenvolvido utilizando a linguagem de programação PHP, JavaScript e banco de dados MySQL. Foi utilizado o software Sublime Text como ambiente de desenvolvimento.

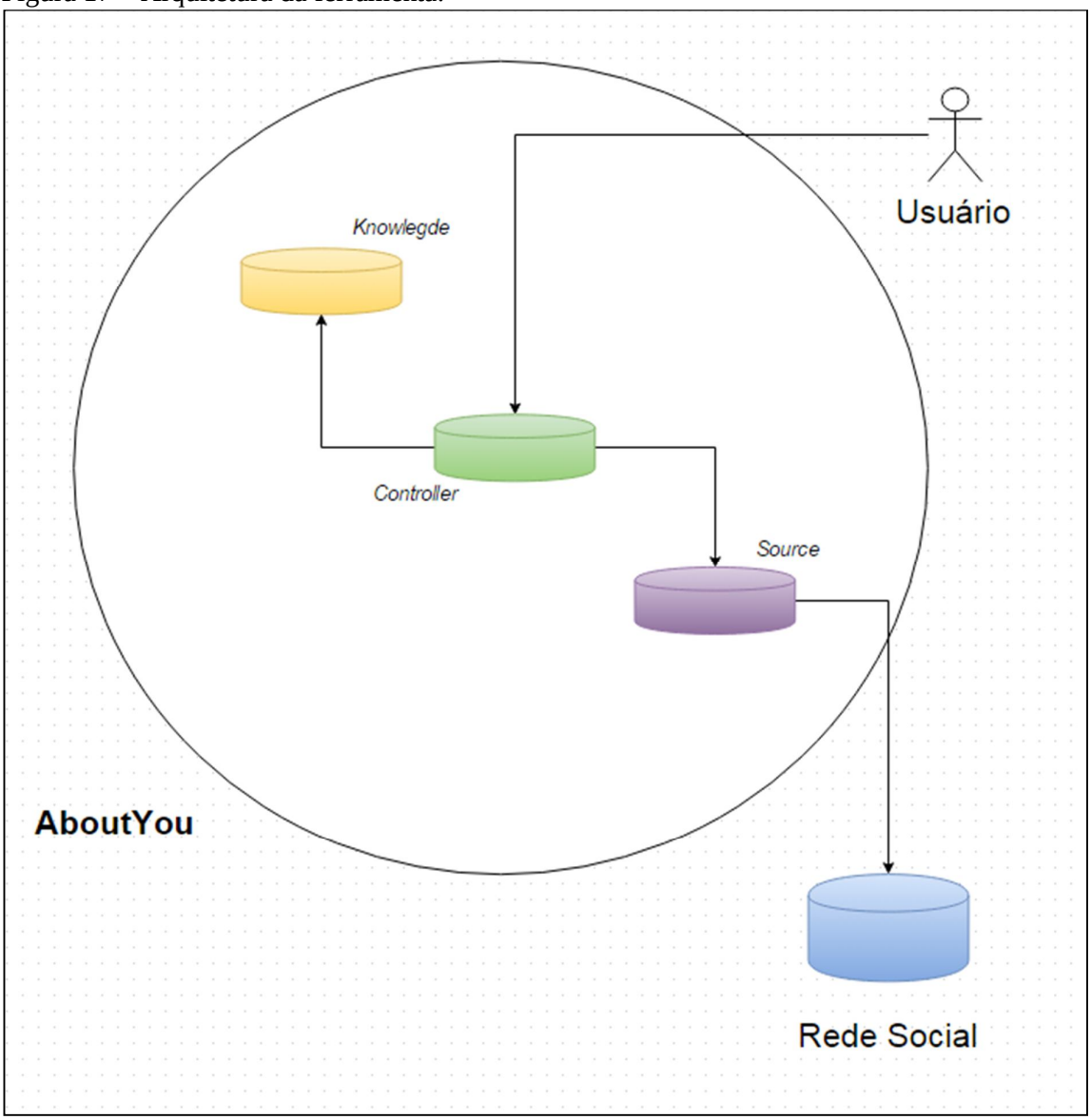
9.2.3 Arquitetura da ferramenta

A arquitetura base do protótipo é constituida em três partes distintas (Figura 17), por conta disso foram separadas em três API's (Interface de Programação de Aplicações), pois dessa forma é possível acessar os dados de qualquer parte da aplicação.

A primeira parte é onde se encontra a ontologia, foi nomeada como knowledge. Essa parte do protótipo tem como objetivo identificar os possíveis perfis do usuário que está sendo pesquisado. O texto fonte é injetado nessa API e ela retornará os perfis, em ordem de prioridade do mais relevante para o menos relevante.

A segunda parte é onde fica a aplicação principal denominada mining que tem por função requisitar os dados do usuário à API source, fazer todos os processos de mineração de dados e por fim quando os texto estiverem o limpo são envialos para a API knowledge. É na API controller que se encontram todas as ligações com as demais API's ou seja, o controlador requisita os dados para a API source fazendo toda a mineração desses dados e por fim envia o texto processado para a API knowledge identificar o perfil.

Figura 17 – Arquitetura da ferramenta.



Fonte: Elaborada pelo autor.

As Figuras 18, 19 e 20 apresentam o layout da aplicação.

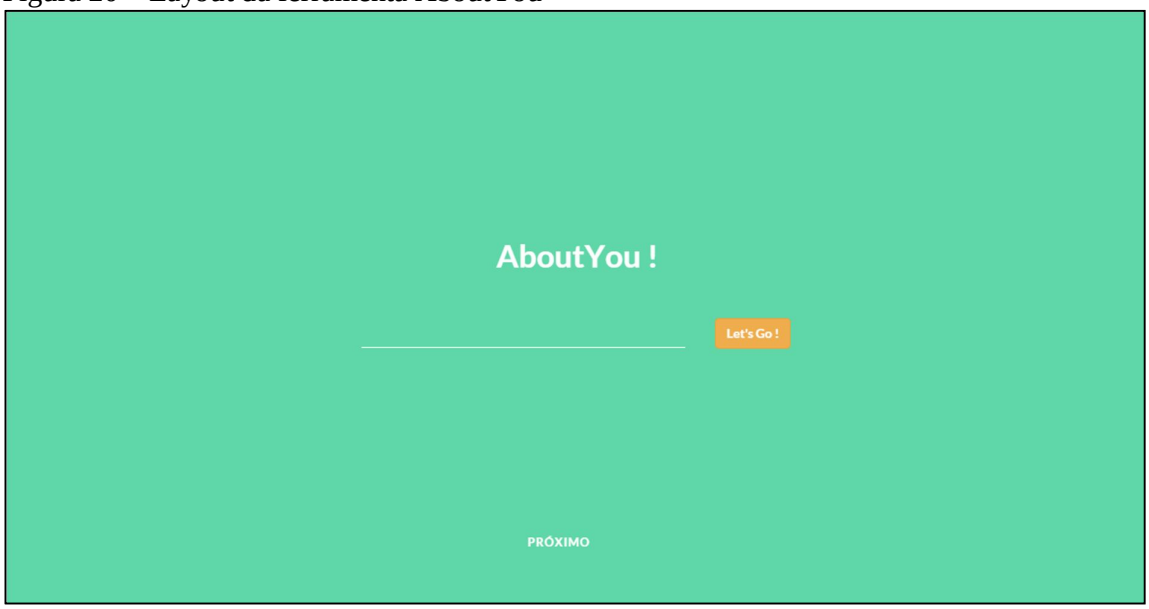
Figura 18 – Layout inicial do AboutYou.

Fonte: Elaborada pelo autor

Figura 19 – Layout de apresentação do autor.

Fonte: Elaborada pelo autor

Figura 20 – Layout da ferramenta AboutYou



Fonte: Elaborada pelo autor.

9.2.4 Seleção dos dados

O primeiro processo na mineração de dados é a seleção de dados. O objetivo desse processo é decidir quais são os conjuntos de dados relevantes.

Para esse primeiro momento foi construído como já citado anteriormente, a API knowledge cuja principal função é requisitar os dados do usuário para obter os últimos dez compartilhamentos que ocorreram dentro do seu Facebook.

Por exemplo se dentre os dez últimos compartilhamentos o usuário que está sendo pesquisado compartilhou um link relacionado a futebol do site Globo Esporte a knowledge receberá esse dado juntamente com os outros nove compartilhamentos que foram requisitadas.

Por questões de segurança a API do próprio Facebook pede para que o perfil que esteja sendo pesquisado autorize a disponibilizar suas informações. Fazendo isso ele cria uma hash (senha) que permite àquela aplicação a requisitar dados daquele perfil. Deste modo, toda vez que é necessário requisitar uma informação desse perfil é necessário, além de enviar o login dele também enviar essa hash para ter autorização na obtenção dos dados.

Assim que esses dados retornam eles vêm em formato JSON, porém não se tem o texto completo da notícia, o que se tem nesse momento é o titulo da matéria e um texto como

subtítulo o que é necessário para identificar qual o assunto do texto, e posteriormente, identificar o perfil do usuário.

9.2.5 Limpeza dos dados

O segundo processo (que já se localiza dentro da API controller) na mineração de dados é a limpeza dos dados em questão retiradas aquelas palavras que podem gerar ruído no resultado. Define-se como ruído um resultado mal estabelecido ou um resultado que possa parecer confuso

Por meio dessa definição foi executado um o pré-processamento do texto que inclui a retirada das chamadas stopwords que correspondem às palavras que em uma busca podem ser consideradas irrelevantes (por exemplo: os, as, de, do,...). Especificamente para esta pesquisa foi elaborada uma lista com cerca de 255 stopwords (a lista completa encontra-se nos anexos deste relatório).

9.2.6 Avaliação do processo

A avaliação tem como objetivo garantir que o modelo gerado atente as expectativas, ou seja, antes de enviar os dados para a knowledge é necessário verificar se os dados processados estão de acordo com o esperado. Para garantir essa qualidade textual foi criado um método se o texto esté conforme o formato estabelecido. Caso não esteja no formato executa-se a fase de limpeza de dados. Se estiver no formato estabelecido é enviado para o próximo passo que será a execução da classificação.

9.2.7 Execução e classificação

Após os passos anteriores a fase de execução é iniciada. Nesta fase é feita basicamente o escalonamento dos conceitos.

Esse processo de escalonamento é feito com o auxilio da ontologia do Yahoo que conforme acusa Fenselet al. (2002), provê uma noção básica de generalização e especialização com a maior parte das relações do tipo "é-um".

Entretanto, nem todas as relações desta ontologia seguem estritamente esse tipo de relação. O processo de generalização proposto é realizado considerando um limite ja que , devido ao grande número de conceitos, a propagação para todos os níveis da ontologia tornaria o processo computacionalmente ineficiente. Do mesmo modo que é realizado no

EXTRATOP, através dessa hierarquização é feita a verificação se palavras presentes no texto correspondem àquelas que descrevem os conceitos ontológicos (mapeamento). O mapeamento realizado pelo AboutYou também manteve as mesmas caracteristicas da forma que é feita no EXTRATOP: é feita a propagação de pesos entre conceitos pais. Assim, sempre que um peso é propagado do conceito-filho para o conceito-pai seu valor é reduzido. No AboutYou, apropagação de pontos para o conceito-pai corresponderá a 50% do peso obtido pelo conceito-filho.

O cálculo do peso de um conceito é feito tendo por base a frequência das palavras no documento que são mapeadas no conceito, ou seja, que correspondem aos descritores dos conceitos. A frequência é identificada pela contagem absoluta das ocorrências da palavra no documento. O cálculo do peso de um conceito, com base na frequência das palavras, parte do princípio de que a repetição de palavras em um texto é feita com o intuito de enfatizar algum assunto e pode ser um indicador de significância das palavras, podendo servir como base para determinar os perfis dos usuários, conforme hipótese deste trabalho. (SALTON, 1983).

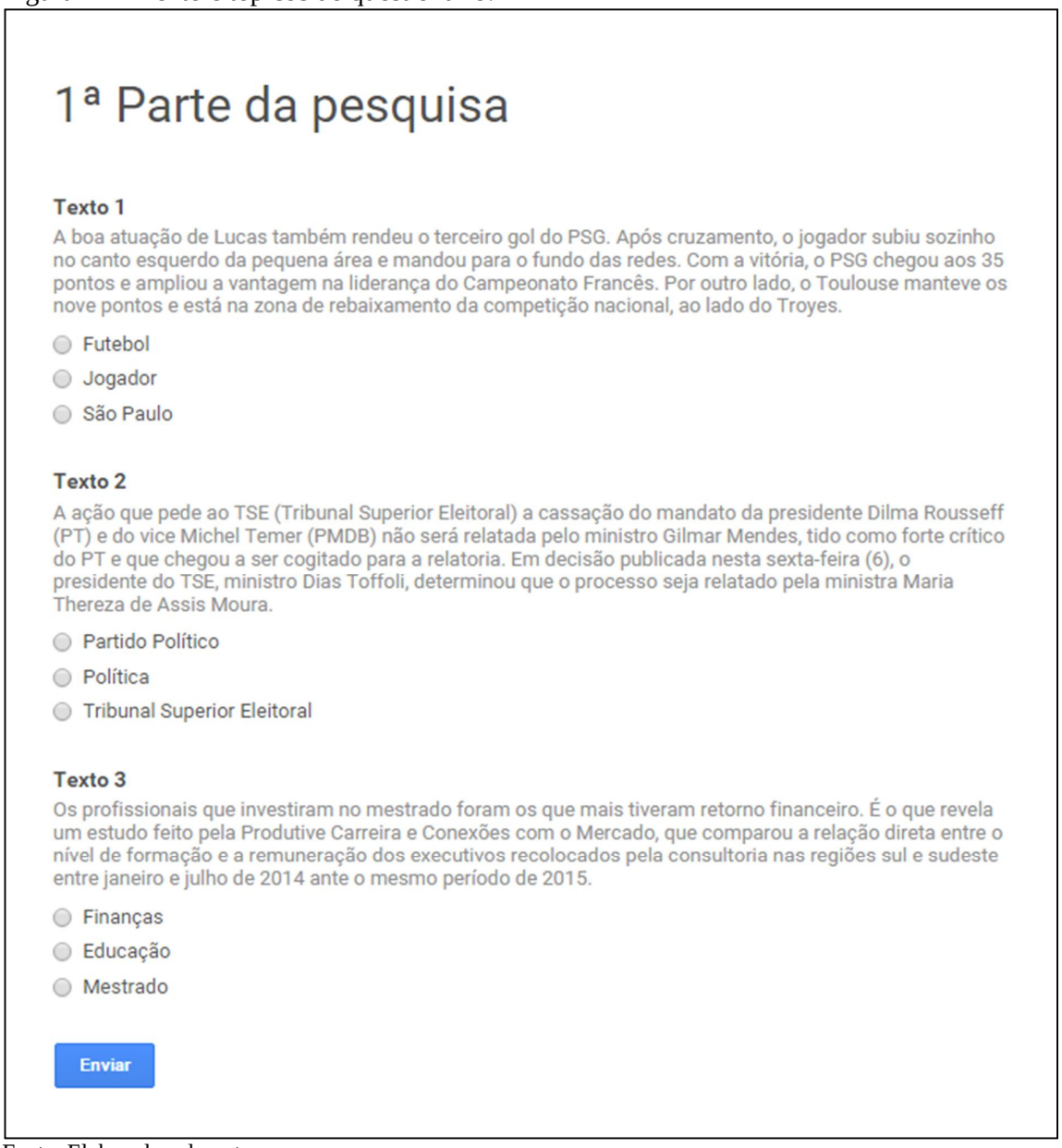
9.2.8 R**esultados da avaliação da ferramenta AboutYou**

Como forma de medir a eficiencia do protótipo foi separado em dois testes.

O primeiro foi relacionado à detecção de tópicos relevantes. O intuito deste teste foi revalidar a ontologa, já que novos descritores foram introduzidos manualmente. Foi elaborado um experimento que envolveu a participação de 15 voluntários. O experimento consistiu em apresentar, para cada voluntário, um conjunto de 3 textos curtos, cada um sobre um tema específico (Futebol, Educação e Política). Associada a cada texto havia uma lista com 6 ou 5 tópicos (definidos automaticamente pelo protótipo como relevantes) que deveriam ser analisados e assinalados caso os voluntários entendessem que os mesmos representassem tópicos relacionados ao textos.

Na Figura 21 é demonstrado os textos utilizados no experimento e os respectivos tópicos associados.

Figura 21 – Texto e tópicos do questionário.



Fonte: Elaborada pelo autor.

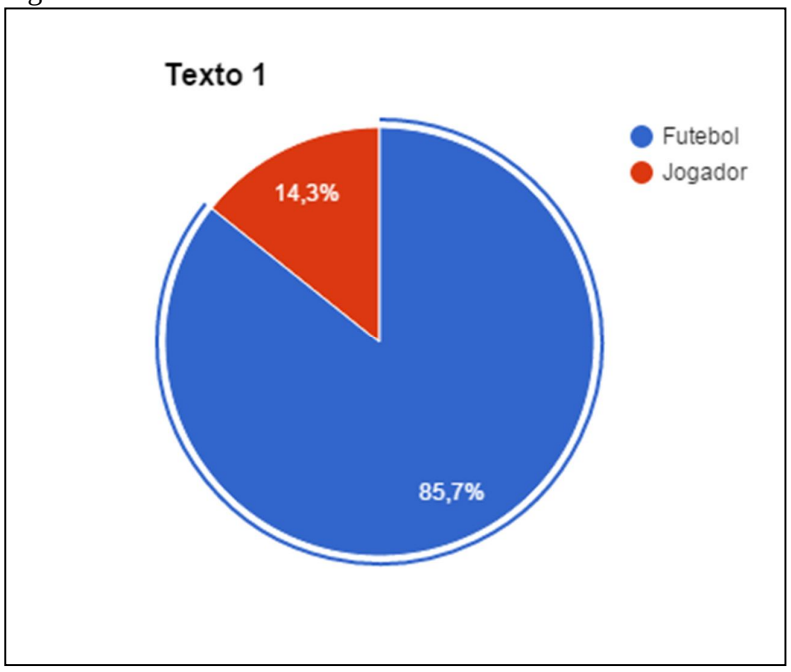
Para avaliar as respostas do protótipo foi verificado, para cada texto, qual o nível de concordância das pessoas com os tópicos apresentados. Apesar de este ser um experimento simples ele é útil para mostrar, em um primeiro momento,a inserção de novos descritores não alteraram a capacidade de a ontologia determinar tópicos relevantes em textos. Para esta análise foi definida como medida de precisão do sistema, a relação entre os 3 tópicos

identificados pelo protótipo para cada texto e aqueles que efetivamente foram referenciados pelos voluntários.

A precisão foi analisada através da porcentagem de concordância entre as respostas dos juízes humanos e os tópicos indicados peo sistema, ou seja, neste sentido quanto maior a porcentagem do tópico maior é a chance de o sistema estar indicando o conceito real do texto.

A Figura 22 mostra os dados coletados referentes ao Texto 1 (Figura 21), relacionado ao tema Futebol. Na figura estão registradas quantas indicações os tópicos receberam. Apesar de serem 3 tópicos apenas os dois abaixo foram pontuados ou seja o último ficou com 0% de votos. Considerando portanto, o primeiro texto pode-se notar que 85,7% das pessoas consideraram Futebol como o tema do texto.

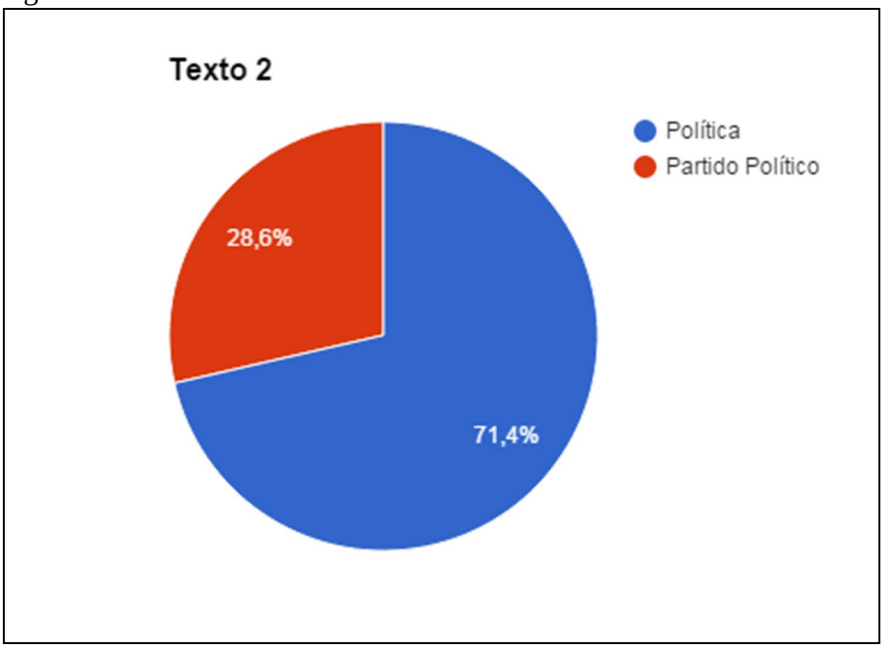
Figura 22 – Gráfico do texto sobre Futebol.



Fonte: Elaborada pelo autor.

A Figura 23 traz os dados referentes ao Texto 2 (Figura 21) sobre Política. De acordo com a figura nota-se que 2 dos 3 tópicos foram apontados também pelos usuários como sendo relevantes. Da mesma forma que o texto anterior o terceiro tópico não foi incluído pois teve 0% dos votos. Porém 71,4% das pesssoas consideraram Políticia o assunto principal.

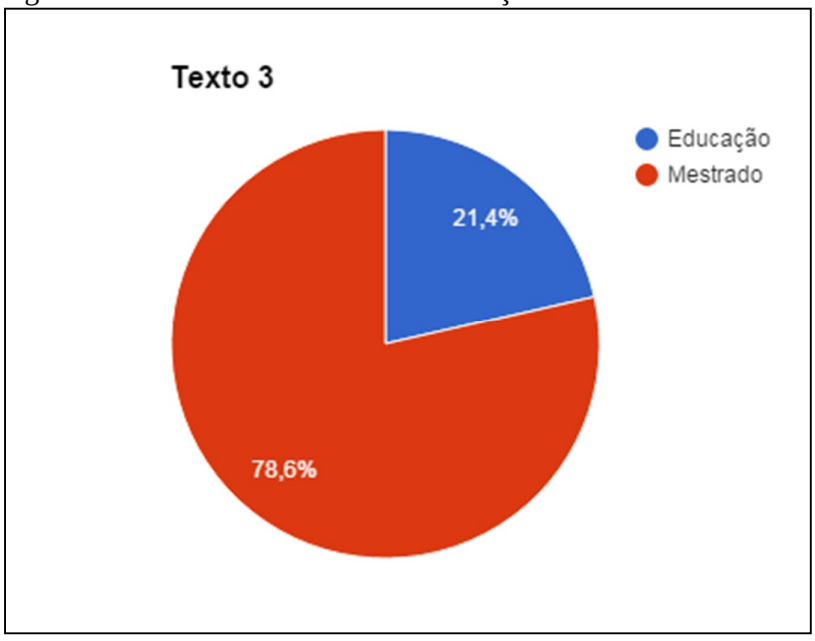
Figura 23 – Gráfico do texto sobre Política.



Fonte: Elaborada pelo autor.

Na análise dos resultados do último texto (Figura 24), 78,6% das pessoas consideraram, o texto referente ao assunto de Mestrado.

Figura 24 – Gráfico do texto sobre Educação.



Fonte: Elaborada pelo autor.

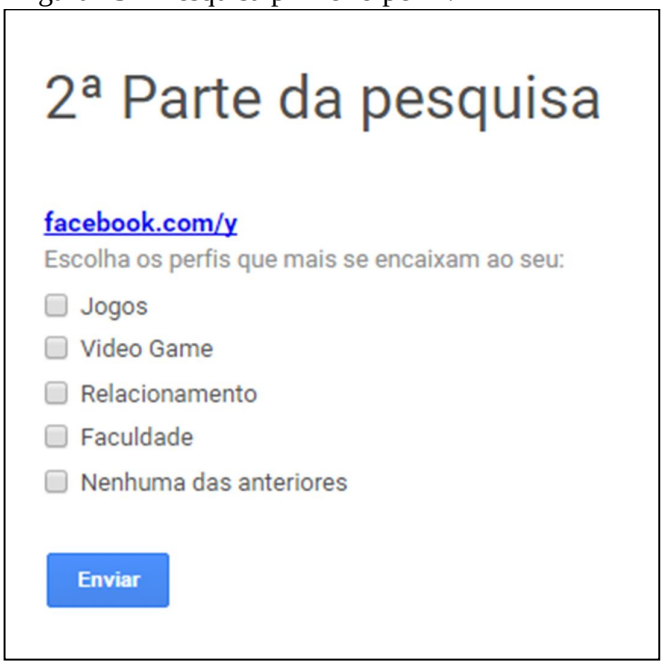
Observa-se que, de um modo geral, os tópicos apontados pelo sistema estão condizentes com o indicado pelos juízes humanos, visto que houe sempre grande concordância, com pelo menos 2 deles.

O segundo teste foi relacionado diretamente com objetivo da pesquisa que é detectar perfis de um usuário através da sua rede social. Todos os testes foram feitos com voluntários. Essa pesquisa selecionou os 4 tópicos relatados pelo protótipo como os perfis do usuário e sendo assim esses tópicos serão apresentados juntamente com um item denominado “Nenhum dos anteriores”, para os caso sem que o usuário não estivessse de acordo de que os itens listados fossem indicativos de seu perfil real.

Os gráficos exibidos a seguir para cada usuário trazem informações sobre todas as opções que foram indicadas pelo protótipo, porém destaca apenas que o usuário marcou como sendo realmente relacionadas ao seu perfil.

O primeiro perfil é relacionado a uma pessoa que gosta muito de jogos de vídeo game e que, com uma frequência alta, compartilha novos jogos ou matérias com videos de jogos.

Figura 25 – Pesquisa primeiro perfil.



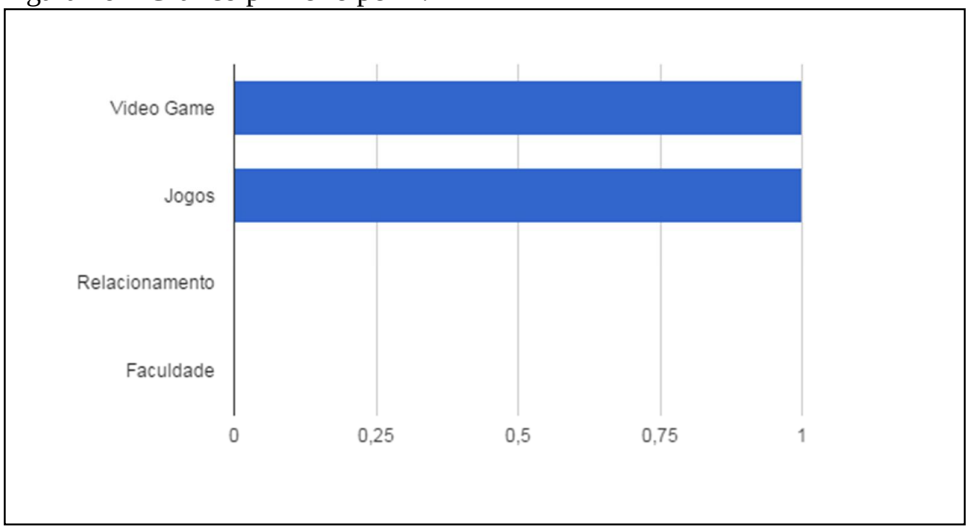
Fonte: Elaborada pelo autor.

Apesar de o princiapal perfil ser um Jogador de Video Game no momento da extração dos dados, essa pessoa também havia postado alguns artigos sobre namoro e também tinha

compartilhado links da faculdade onde ele estuda, o que justifica a ferramenta ter indicado os outros tópicos.

No relacionado a esse usuário nota-se que os principais perfis considerados pelo protótipo (os que aparecem nas primeiras posições da lista) também foram os indicados pelo próprio usuário. No caso Jogos e Video Game, foram os perfis indicados pelo usuário.

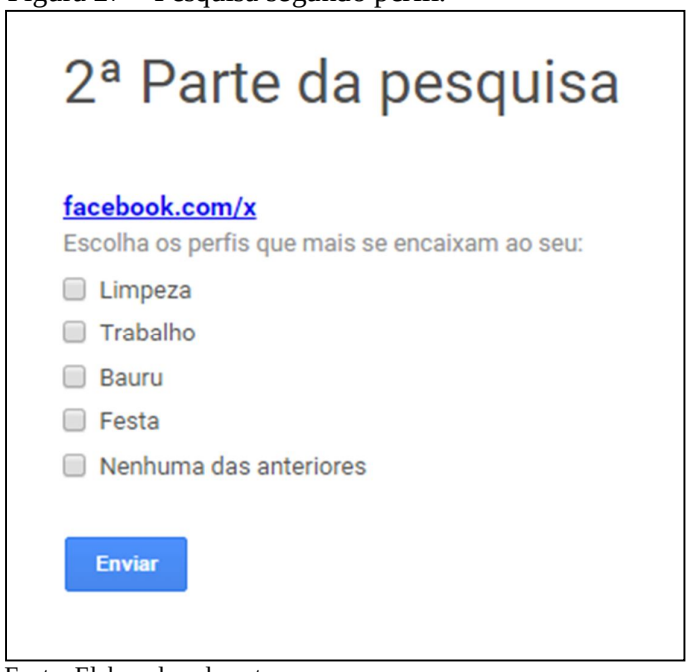
Figura 26 – Gráfico primeiro perfil.



Fonte: Elaborada pelo autor.

O segundo era um perfil de uma pessoa que trabalhava em uma empresa de limpeza de casas domésticas e em sua rede social ela compartilhava vários links do seu trabalho.O interessante desse perfil é que a empresa focava em serviços para casas que continham estudantes de faculdade. Isso justifica o tópico Festa dentre os outros, pois a empresa trabalhava o seu marketing procurando pessoas que queriam uma limpeza após uma festa. Nesse caso os perfis mais relevantes indicados pela ferramenta foram Limpeza e Trabalho.

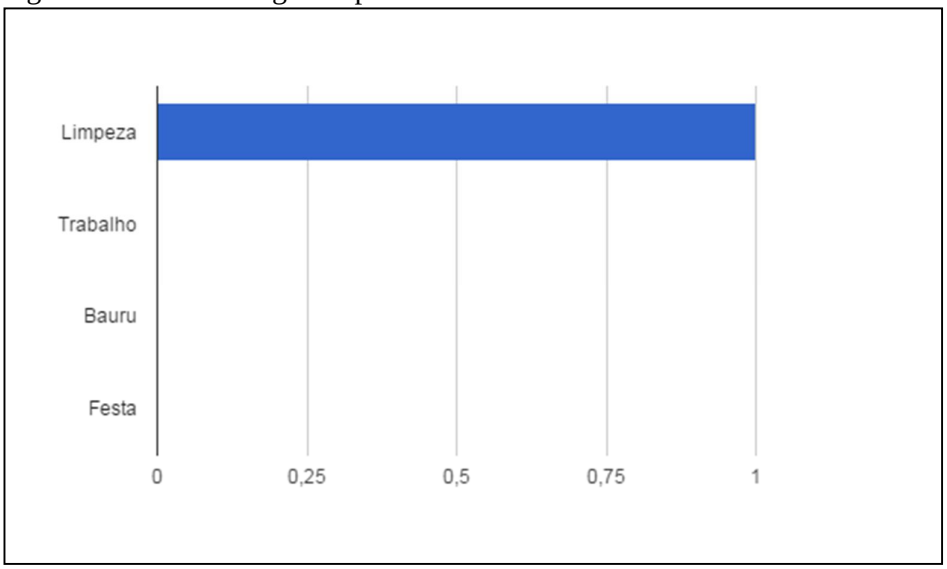
Figura 27 – Pesquisa segundo perfil.



Fonte: Elaborada pelo autor.

O gráfico obtido desse segundo perfil foi o único que aparece marcado apenas um tópico, no caso marcou Limpeza (Figura 28).

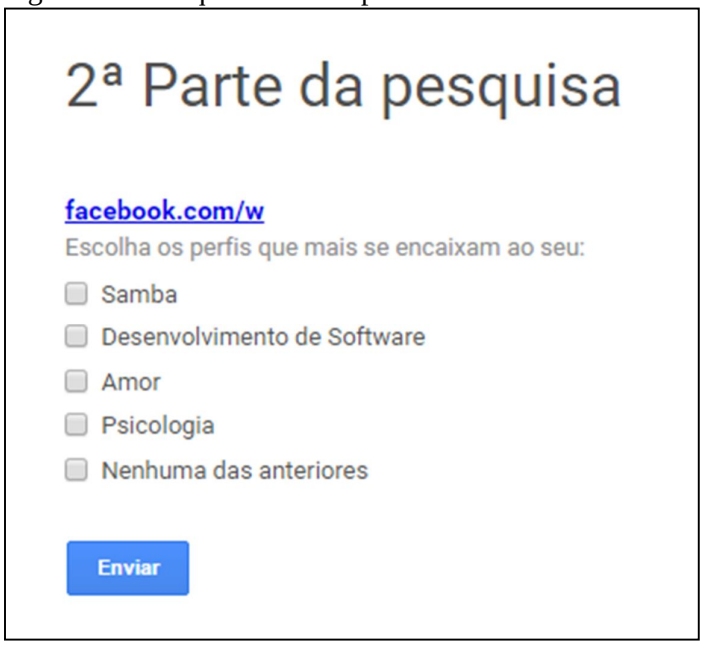
Figura 28 – Gráfico segundo perfil.



Fonte: Elaborada pelo autor.

O último perfil teve itens muitos distintos pois compartilhava links variados na sua rede social. Entretanto, conforme padrão da ferramenta, somente os 4 tópicos mais relevanres foram considerados (Figura 29).

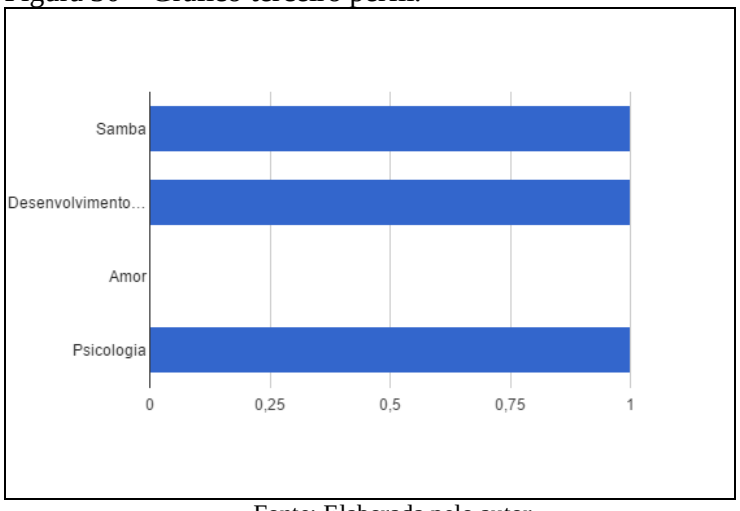
Figura 29 – Pesquisa terceiro perfil.



Fonte: Elaborada pelo autor.

Como se trata de um perfil peculiar é natural que o mesmo marque mais itens que os demais, pois não existe apenas um assunto focado que ele se identifique, porém mesmo assim o perfil não considerou a opção “Nenhuma das anteriores” (Figura 30).

Figura 30 – Gráfico terceiro perfil.



Fonte: Elaborada pelo autor.

10 CONSIDERAÇÕES FINAIS

Numa análise geral de todos os resultados, verifica-se o protótipo obteve uma precisão média de 73,4%, quando se refere à detecção correta de tópicos com base no conteúdo dos documentos (conforme testes preliminares do EXTRATOP). Mesmo no processo de revalidação da ontologia, após sua atualização com a inserção de novos descritores, nos três testes realizados, houve concordância de todos os juízes de que os dois tópicos apontados como principais pelo sistema eram, de fato, relevantes. Apenas o terceiro tópico aparenta não ter uma relação tão relevante com o perfil, o que justifica somente considerar os dois primeiros no momento de traçar o perfil dos usuários. Esse resultado é considerado satisfatório, pois, de um modo geral, a ferramenta aponta para tópicos que estão, de fato, relacionados aos textos e com as postagens nas redes sociais; o que pode indicar potencialidades na abordagem sugerida nesta investigação. Apesar de ter passado por uma etapa de atualização em relação à sua primeira versão (para ser usada no AboutYou), muitos ajustes com relação ao conteúdo da ontologia e mesmo com relação à forma de pontuação de tópicos devem ser feitos já que, pelo modo que a mesma foi estruturada, quando se considera um conjunto maior (acima de 2), os tópicos associados pela ferramenta aparentemente destoam do perfil do usuário. Entretanto, observa-se que, de um modo geral, os tópicos apontados pelo sistema AboutYou quando se considera apenas os dois tópicos mais relevantes para caracterizar o perfil dos usuários, estão condizentes com o indicado pelos juízes humanos, visto que houve sempre grande concordância, conforme resultados dos testes.

As dificuldades nesse processo ocorrem devido ao fato de a língua portuguesa possuir várias palavras com mais de um sentido, gerando ambiguidade e, por lidar apenas de modo superficial com a linguagem, a ferramenta não faz nenhum tipo de desambiguação. Por outro lado, observa-se também que existem palavras presentes na ontologia que, por serem bem específicas de certos temas, ajudam a ferramenta a definir claramente os tópicos contidos nos textos. Isso gera uma diferença de resultados considerando-se as diferentes temáticas dos textos. Por exemplo, no experimento realizado os textos referentes à “Educação” e “Política” foram os que apresentaram a maior precisão. Isso é ocorre por dois motivos possíveis: primeiro, existem muitas palavras específicas que se referenciam diretamente e unicamente aos temas abordados ou então, os temas citados tem mais palavra associadas na ontologia, até mesmo porque não houve nenhum tipo de preocupação em balancear o número de termos associados a cada conceito.

REFERÊNCIAS

AGIRRE, E. et al. **Enriching WordNet concepts with topic signatures**. In: SIGLEX WORKSHOP ON WORDNET AND OTHER LEXICAL RESOURCES: APPLICATIONS, EXTENSIONS AND CUSTOMIZATIONS, 2001, Pittsburg. Proceedings…Cambridge: MIT Press, 2001. p. 1-7.

AMORIM T. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados.** Universidade Federal de Pernambuco, 2006.

BERRY, M. J. A.; LINOFF, G. **Data Mining Tehniques – for marketing, sales, andcustomer support**.United States: Wiley Computer Publishing, 1997.

BITTENCOURT, Guilherme. **Brevehistória da Inteligência Artificial**. 1999.

CAMILO O. C.; SILVA C. J. **Mineração de dados: Conceitos, Tarefas, Métodos e Ferramentas.** Universidade Federal de Goiás, 2009.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R.**What areontologies, and why do we need them?** IEEE Intelligent Systems, v. 14, n. 1, p. 20-26,1999.

CHRISTOPHI, C. **Mining keywords from large topic taxonomies**. 2004. Dissertação(Mestrado) - Department of Computer Science, University of Cyprus, Nicosia, 2004.

FERNANDES, A. M. da R. **Inteligência Artificial**: noções gerais. 2. ed. Visualbooks Florianopolis, 2005.

FREITAS, F. L. G. D. **Ontologias e a Web semântica**. In: Congresso da Sociedade Brasileira de Computação, 23., 2003, Campinas. Anais...Campinas: SBC, 2003. p. 1-52.

GENESERETH, R. M.; NILSSON, L. **Logical foundarions of AI**. Los Altos: Morgan Kaufman, 1987.

GOMEZ-PÉREZ,A., **Tutorial onOntologicalEngineering, Internacional Joint Conferenceon Artificial Intelligence** – IJCAI, 1999.

GRUBER, T. R. A translation approach to portableontologies**.KnowledgeAcquisition**,v. 5, n. 2, p. 199-220, 1993.

GUARINO, N.; CARRARA, M.; GIARETTA, P. Formalizingontologicalcommitment. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 12., 1994, Seattle. **Proceedings**… San Fransisco: Morgan Kaufmann, 1994. p. 560-567.

ISAÍAS, Pedro; Carvalho, Tiago; Assis, Ana Cristina (1995), **Informação Multimédia na Internet** (Braga, 6-8 de Julho de 1995).

JUNIOR, J. B. dos S**. Introdução à Inteligência Artificial**. 1999. Universidade de São Paulo - USP, São Paulo, 1999.

LAROUSSE. **Grande Enciclopédia Larousse Cultural. Editora Nova Cultural**, 1999.

LEITE, D.S.; RINO, L.H.M. **Selecting a Feature Set toSummarizeTextsin BrazilianPortuguese**. In: INTERNATIONAL JOINT CONFERENCE IBERAMIA/SBIA, 2006,RibeirãoPreto. Proceedings... Heildelberg : Springer-Verlag.

LIN, C. **Knowledge-based automatic topic identification**. In: ANNUAL MEETING ONASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 33., 1995, Morristown.Proceedings... Cambridge: MIT Press, 2004. p. 308-310.

LOH, S. **Uma abordagem baseada em conceitos para descoberta de conhecimento em textos.** 2001. 110 f. Tese (Doutorado) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001.

MAEDCHE; A., STAAB; S. "**Measuring similarity between ontologies**". In: ProceedingsoftheEuropeanConferenceon EKAW, 2002.

MILLER, G.A**. WordNet: a lexical database for english**. Communications of the ACM,v. 38, n. 11, p. 39-41, 1995.

MLADENIC, D.; GROBELNIK, M. **Assigning keywords to documents using machine learning.** In: INTERNATIONAL CONFERENCE ON INFORMATION AND INTELLIGENT SYSTEMS,10., 1999, Varazdin. Proceedings… Varazdin: Faculty of Organization and Informatics, University of Zagreb 1999. p. 123-131.

NASCIMENTO, T. **A importância dos protótipos no desenvolvimento de sistemas.** 2013. Bootstrap. Disponivel em: < http://thiagonasc.com/desenvolvimento-web/a-importancia-dos-prototipos-no-desenvolvimento-de-sistemas >. Acesso em 14 de dezembro de 2015.

NIKOLOPOULOS, Chris. Expert Systems – **Introduction to First and Second Generation and Hybrid Knowledge Based Systems**. Marcel Dekker Inc. Press.1997.

NOVELLO, T. C. **" Ontologias, sistemas baseados em conhecimento e modelos de banco de dados**," Universidade Federal do Rio Grande do Sul, 2002.

PALMER, M. Miltilingualresources. **LinguisticaComputazionale**, v.14-15, 2001.

PARDO, T. A. S.; RINO, L. H. M. **TeMário: um corpus para sumarização automáticade textos.** São Carlos-SP: USP/ICM, 2003. Série de Relatórios do NILC; NILC-TR-03-09.

PEDREIRA-SILVA, P. ; FREITAS, C. ; SILVA, E. G. **Desenvolvimento de um software para detecção automática de tópicos em documentos textuais baseada em taxonomia**. In:

Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2014, São Carlos - SP. BDBCompProceedings, 2014.

PEDREIRA-SILVA, P. **ExtraWeb: um sumarizador de documentos Web baseado em etiquetas HTML e ontologia.** Dissertação (Mestrado) – Departamento de CiênciadaComputação , Universidade Federal de São Carlos, 2006.

PRESSMAN, Roger S**. Software Engineering: a Practitioner's Approach.** 6. ed. McGraw-Hill, 2005

RUSSEL, S.; NORVIG, P. **Inteligência artificial**. Tradução: Stuart Russel, Peter Norvig. Rio de Janeiro: Elsevier, 2011.

SALTON, G. ; **Introduction to Modern Information Retrieval.** 1983. Mcgraw-Hill College, 1983.

SANTOS L. C. B. **Aprendizagem, cognição e Inteligência Artificial.** 2006. Universidade Estadual de Campinas (Unicamp), Campinas. 6f.
Disponível em < http://www.dca.fee.unicamp.br/~gudwin/courses/IA889/2006/IA889-02.pdf>

TIUN, S.; ABDULLAH, R.; KONG, T. E. **Automatic topic identification usingontology hierarchy.** In: CICLING: CONFERENCE ON INTELLIGENT TEXTPROCESSING AND COMPUTATIONAL LINGUISTICS, 2., 2001, Mexico City.Proceedings... Heildelberg : Springer-Verlag, 2001. p. 444-453.

VOSSEN, P. EuroWordNet: Linguistic ontologies in a multilingual database. **Communication and Cognition for Artificial Intelligence** (Special Issue), v. 15, n. (1-2), p. 37-80, 1998a.

WIVES, L.K. **Utilizando conceitos como descritores de textos para o processo deidentificação de conglomerados (clustering) de documentos**. 2004. 136 f. Tese(Doutorado) - Instituto de Informática, Universidade Federal do Rio Grande do Sul,Porto Alegre, 2004.

WU, C. W.; LIU, C. L. **Ontology-based text summarization for business news articles**. In: ISCA: INTERNATIONAL CONFERENCE ON COMPUTERS ANDTHEIR APPLICATIONS, 18., 2003, Honolulu.Proceedings… Cary: ISCA, 2003. p. 389-392.

ANEXO

Stop List

a
à
agora
ainda
alguém
algum
alguma
algumas
alguns
ampla
amplas
amplo
amplos
ante
antes
ao
aos
após
aquela
aquelas
aquele
aqueles
aquilo
as
até
através
cada
coisa
coisas
com
como
contra
contudo
da
daquele
daqueles
das
de
dela
delas
dele
deles
depois
dessa
dessas
desse
desses
desta
destas
deste
deste
destes
deve
devem
devendo
dever
deverá
deverão
deveria
deveriam
devia
deviam
disse
disso
disto
dito
diz
dizem
do
dos
e
é
e'
ela
elas
ele
eles
em
enquanto
entre
era
essa
essas
esse
esses
esta
está
estamos
estão
estas
estava
estavam
estávamos
este
estes
estou
eu
fazendo
fazer
feita
feitas
feito
feitos
foi
for
foram
fosse
fossem
grande
grandes
há
isso
isto
já
la
la
lá
lhe
lhes
lo
mas
me
mesma
mesmas
mesmo
mesmos
meu
meus
minha
minhas
muita
muitas
muito
muitos
na
não
nas
nem
nenhum
nessa
nessas

nesta
nestas
ninguém
no
nos
nós
nossa
nossas
nosso
nossos
num
numa
nunca
o
os
ou
outra
outras
outro
outros
para
pela
pelas
pelo
pelos
pequena
pequenas
pequeno
pequenos
per
perante
pode
pôde
podendo
poder
poderia
poderiam
podia
podiam
pois
por
porém
porque
posso
pouca
poucas
pouco
poucos
primeiro
primeiros
própria
próprias
próprio
próprios
quais
qual
quando
quanto
quantos
que
quem
são
se
seja
sejam
sem
sempre
sendo
será
serão
seu
seus
si
sido
só
sob
sobre
sua
suas
talvez
também
tampouco
te
tem
tendo
tenha
ter
teu
teus
ti
tido
tinha
tinham
toda
todas
todavia
todo
todos
tu
tua
tuas
tudo
última
últimas
último
últimos
um
uma
umas
uns
vendo
ver
vez
vindo
vir
vos
vós