

Dissertação de Mestrado

Construção e Refinamento de Perfil Baseado em Dados Georreferenciados e de Deslocamento

João Pedro Pontes Lima Matias

joaopedro.plima@gmail.com

Orientadores:

Prof. Dr. Evandro de Barros Costa

Prof. Dr. Patrick Henrique da Silva Brito

Maceió, Novembro de 2012

João Pedro Pontes Lima Matias

Construção e Refinamento de Perfil Baseado em Dados Georreferenciados e de Deslocamento

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Modelagem Computacional de Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientadores:

Prof. Dr. Evandro de Barros Costa

Prof. Dr. Patrick Henrique da Silva Brito

Maceió, Novembro de 2012

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Modelagem Computacional de Conhecimento do Instituto de Computação da Universidade Federal de Alagoas, aprovada pela comissão examinadora que abaixo assina.

Prof. Dr. Evandro de Barros Costa - Orientador
Instituto de Computação
Universidade Federal de Alagoas

Prof. Dr. Patrick Henrique da Silva Brito - Orientador
Instituto de Computação
Universidade Federal de Alagoas

Prof. Dr. Hyggo Oliveira de Almeida - Examinador
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande

Prof. Dr. Henrique Pacca Loureiro Luna - Examinador
Instituto de Computação
Universidade Federal de Alagoas

Prof. Dr. Márcio de Medeiros Ribeiro - Examinador
Instituto de Computação
Universidade Federal de Alagoas

*Com bastante amor, dedico esta dissertação especialmente
à minha mãe. A pessoa mais guerreira e batalhadora que
eu poderia conhecer na vida.*

Agradecimentos

Agradeço primeiramente a Deus pela vida, por todas as bênçãos e por ter me fornecido todas as ferramentas necessárias para chegar até este momento. Através da sua sabedoria, colocou barreiras e desafios em meu caminho, para poder evoluir e conseguir concluir mais esta etapa.

À minha família, em especial à minha mãe, que sempre esteve comigo, compartilhando todas as expectativas, alegrias, tristezas, ansiedades e conquistas ao longo do caminho, sempre na torcida para que tudo desse certo. Além disso, o seu constante apoio, estímulo, amor e carinho, me forneceram todo o suporte necessário para seguir em frente.

À minha amiga e companheira Carolina por, mesmo à distância, ter me dado a força, o estímulo e o suporte necessários para percorrer mais esta jornada, além das palavras de amor e carinho constantes. Agradeço pelas alegrias que ela tem trazido em nosso relacionamento.

A todos os companheiros da ZUQ/LITS, em especial ao Prof. Willy Tiengo e ao Márcio Aguiar (vulgo Bisonho), que sempre proporcionaram um ambiente de trabalho descontraído e de constante aprendizado, e contribuíram constantemente com várias sugestões. Através das inúmeras discussões sobre os mais variados temas, deram sua contribuição para meu amadurecimento pessoal e profissional. Grande parte deste trabalho foi desenvolvida graças ao laboratório, que disponibilizou uma ótima estrutura e um ambiente bastante motivador.

Aos professores e servidores do Instituto de Computação e aos membros da banca, em especial aos meus orientadores Prof. Evandro Costa e Prof. Patrick Brito por todo direcionamento, apoio e ensinamentos. Agradeço também ao Prof. Márcio Ribeiro, que com paciência e presteza me ajudou a fazer todas as correções necessárias e à Profª. Silvana Calheiros do Instituto de Geografia, Desenvolvimento e Meio Ambiente desta universidade, que contribuiu, de modo fundamental, com os dados georreferenciados existentes no presente trabalho.

A todos os meus amigos, em particular à Rayssa Lira, Everton Cleiton, David Lima, Tamer Cavalcante e Dieges Lima pelo grande auxílio na execução das entrevistas do estudo de caso. Sem eles, não seria possível realizar tantas entrevistas em tempo hábil.

À Fundação de Amparo à Pesquisa do Estado de Alagoas, que sempre forneceu o suporte financeiro necessário para o desenvolvimento desta pesquisa.

Enfim, fico grato a todos que contribuíram comigo na construção deste trabalho.

João Pedro Pontes

“Por vezes, sentimos que aquilo que fazemos não é senão uma gota de água no mar. Mas o mar seria menor se lhe faltasse uma gota.”

Madre Teresa de Calcutá

Resumo

A utilização de dispositivos móveis vem crescendo bastante mundialmente, principalmente como meio de buscar/difundir informação. E com este crescimento, começaram a surgir serviços que buscam facilitar o nosso dia-a-dia. Um que tem se destacado é o serviço baseado em localização ou *location-based services* (LBS). Serviços deste tipo buscam trazer informações que sejam sensíveis à localização atual do usuário, como informações turísticas, restaurantes próximos ou áreas de congestionamentos. Dentre os vários tipos de LBS, podemos destacar o que envia anúncios baseados em localização ou *location-based advertising* (LBA), foco do presente trabalho. Desta forma, este trabalho apresenta uma abordagem que, baseada em dados georreferenciados e de deslocamento de pessoas, propõe a construção e refinamento de perfis automaticamente, para posteriormente recomendar anúncios publicitários mais adequados a estes perfis.

Palavras-chave: *Location-based advertising*. Dados georreferenciados. Refinamento de perfis.

Abstract

The use of mobile devices has been quickly growing, mainly as a mean of seeking/disseminating information. With this growth, services that aim to facilitate our routine began to emerge. One that stands out is the location-based services (LBS). These kind of services try to gather information related to the user's current location, such as tourist information, nearby restaurants or traffic jam area. Among the various types of LBS, we emphasize advertisements based on location, called location-based advertising (LBA), focus of the current work. Thus, this work presents an approach, based on geo-referenced data and people movement, that proposes the construction and refinement of profiles automatically, and then recommending the most appropriate advertisements to these profiles.

Keywords: Location-based advertising. Geocoded data. Profile refinement.

Lista de Equações

2.1	Definição da métrica de similaridade D	29
2.2	Representação do cosseno de Salton	29
2.3	Representação do cosseno de Salton em termos de vetores unitários	30

Lista de Figuras

1.1	Diagrama que apresenta a ideia geral da proposta do trabalho	18
2.1	Tipos de <i>Location-based Advertising</i>	23
2.2	Modelo típico de uma rede bayesiana	26
2.3	Árvore de decisão que indica se uma pessoa deve pegar um ônibus	28
2.4	A similaridade por cosseno	30
2.5	O método GQM	31
3.1	Núcleo da ontologia do LASA	34
3.2	Classe Produtos seguindo a estrutura p-Tree	35
4.1	Diagrama de atividades do perfil	40
4.2	Diagrama de atividades da recomendação	40
4.3	Exemplo de um histórico de localização	41
4.4	Histórico de localização e Base de Dados Georreferenciada	43
4.5	Comunicação do Integrador com os módulos de regras e da rede bayesiana	44
4.6	Modelo da Rede Bayesiana compilada	46
4.7	Exemplo de algumas regras da Base de Regras	47
4.8	Comparação do perfil probabilístico com os demais perfis	48
4.9	Histórico de localização obtido pelo uso do sistema	50
4.10	Exemplo de uma tabela de probabilidade condicional para o atributo gênero	51
4.11	Exemplo de uma rede bayesiana compilada	52
4.12	Regras que são usadas para inferir a ocupação para o exemplo	52
4.13	Rede bayesiana com a evidência de ocupação inserida	53
4.14	Classe que instancia e alimenta a rede bayesiana	55
4.15	Classe que gerencia e executa a base de regras	56
4.16	Exemplo de algumas regras da Base de Regras	57
4.17	Método que compara o público-alvo do anunciante com o usuário (Parte 1)	58
4.18	Método que compara o público-alvo do anunciante com o usuário (Parte 2)	59
5.1	Mapa que indica os locais das entrevistas	61
6.1	Gráfico da distribuição de idade	66
6.2	Gráfico da distribuição de gênero	66
6.3	Gráfico da distribuição de escolaridade	67
6.4	Gráfico da distribuição de renda	67
6.5	Gráfico da distribuição de ocupação	68
6.6	Gráfico da frequência média de uso dos ônibus	68
6.7	Resultado dos testes das regras da Base de Regras	70
6.8	Gráfico do teste final para a concessionária	74
6.9	Gráfico do teste final para a farmácia	75
6.10	Gráfico do teste final para o centro de estética	75
6.11	Gráfico do teste final para a auto escola	76
6.12	Gráfico do teste final para a empresa de TI	76

Lista de Tabelas

1.1	Principais características do trabalho proposto	16
2.1	Classificação dos serviços baseados em localização	22
2.2	Push vs Pull	24
4.1	Comparação Faixa Etária X Ocupação	44
4.2	Tabela de públicos-alvo dos anunciantes	51
6.1	Valores do teste final para o público-alvo da concessionária de veículos	72
6.2	Valores do teste final para o público-alvo da farmácia	72
6.3	Valores do teste final para o público-alvo do centro de estética	73
6.4	Valores do teste final para o público-alvo da auto escola	73
6.5	Valores do teste final para o público-alvo da empresa de TI	73
C.1	Valores do pré-teste para o público-alvo da concessionária de veículos	88
C.2	Valores do pré-teste para o público-alvo da farmácia	88
C.3	Valores do pré-teste para o público-alvo do centro de estética	88
C.4	Valores do pré-teste para o público-alvo da auto escola	89
C.5	Valores do pré-teste para o público-alvo da empresa de TI	89
C.6	Valores do teste final para o público-alvo da concessionária de veículos	89
C.7	Valores do teste final para o público-alvo da farmácia	89
C.8	Valores do teste final para o público-alvo do centro de estética	90
C.9	Valores do teste final para o público-alvo da auto escola	90
C.10	Valores do teste final para o público-alvo da empresa de TI	90

Sumário

1 INTRODUÇÃO	14
1.1 Contextualização	16
1.2 Objetivos	17
1.2.1 Objetivo Principal	17
1.2.2 Objetivos Específicos	17
1.3 Visão geral da proposta	18
1.4 Estrutura da dissertação	19
2 FUNDAMENTAÇÃO TEÓRICA	20
2.1 Location-based Services	20
2.1.1 Location-based Advertising	21
2.2 Redes Bayesianas	24
2.3 Árvores de Decisão	27
2.4 Métricas de Similaridade	29
2.4.1 Similaridade por cosseno	29
2.5 O método Goal/Question/Metric	31
3 TRABALHOS RELACIONADOS	33
3.1 Estado da Arte	33
3.2 Discussão	39
4 MODELO DE CRIAÇÃO DO PERFIL	40
4.1 Workflow da solução	40
4.2 Definir perfil inicial do usuário	41
4.3 Inferir ocupação do usuário	45
4.4 Refinar os demais campos do perfil	47
4.5 Comparar os perfis dos usuários com os públicos-alvo dos anunciantes	47
4.6 Classificar em ordem de similaridade	49
4.7 Recomendar o anúncio mais apropriado a cada perfil de usuário	49
4.8 Exemplo passo-a-passo	49
4.9 Limitações	54
4.10 Implementação da proposta	54
5 DESCRIÇÃO DO ESTUDO DE CASO	60
5.1 Descrição geral da metodologia utilizada	60
5.2 Estruturação segundo o método GQM	62
5.2.1 Goal	62
5.2.2 Questions	63
5.2.3 Metrics	63
5.3 Definição da avaliação dos resultados	64

6 ANÁLISE DOS RESULTADOS	65
6.1 Perfil dos entrevistados	65
6.2 Público-alvo dos anunciantes	69
6.3 Métrica para G1	70
6.4 Métricas para G2	71
6.5 Discussão dos resultados	73
7 CONCLUSÃO E TRABALHOS FUTUROS	77
7.1 Principais contribuições da pesquisa	77
7.2 Trabalhos Futuros	78
Referências	79
Apêndice A - Questionário para identificação de perfis - Pré-teste	83
Apêndice B - Questionário para identificação de perfis - Teste final	85
Apêndice C - Resultados Completos do estudo de caso	88

1 Introdução

Nos últimos anos, observou-se a disseminação do uso de dispositivos móveis em todo mundo [Liapis et al. \(2008\)](#), incluindo nas classes de baixo poder aquisitivo [IDGNow! \(2011\)](#). Os *smartphones* estão se tornando cada vez mais comuns como meio de difundir e/ou buscar informação em larga escala, principalmente entre os que estudam e trabalham. Um estudo aponta que as vendas de *smartphones* para usuários finais atingiram 115 milhões de unidades no terceiro trimestre de 2011, um incremento de 42% em relação ao mesmo período de 2010 [ComputerWorld \(2011\)](#). Segundo [Jordan \(2012\)](#), no Brasil foram habilitadas aproximadamente 39 milhões de novas linhas, uma alta de 19%. Isso se deve basicamente ao barateamento e popularização dos dispositivos móveis, além do grande crescimento da infraestrutura de telecomunicações e, em parte, à implantação do 4G e expansão do 3G.

Em 2011, pela primeira vez na história, os fabricantes de *smartphones* entregaram mais produtos que os fabricantes de computadores pessoais (*tablets, notebooks, netbooks e desktops*) [Service \(2012\)](#). A partir disso, vários estudos vêm sendo realizados para usos e aplicações nessa área. Nesse contexto, um tema que tem ganhado bastante relevância é o de serviços baseados em localização – *Location-based Services (LBS)*. De acordo com [Kölmel & Spiros \(2002\)](#), *location-based services* são serviços que exploram o conhecimento sobre o lugar onde o usuário de um dispositivo móvel está localizado, utilizando para isso, várias tecnologias para obtenção da localização geográfica. [Liapis et al. \(2008\)](#) cita que alertas sobre congestionamentos, informações turísticas e publicidade personalizada são exemplos típicos deste tipo de serviço de informação sensível ao contexto.

Um tópico importante dentro de LBS é o de publicidade baseada em localização – *Location-based Advertising (LBA)* [Liapis et al. \(2008\)](#). Muitos consideram que os dispositivos móveis são a forma mais eficaz de realizar uma publicidade, seja de forma direta ou indireta [Marketer \(2009\)](#). Além disso, nesses últimos anos, grandes investimentos tem sido feitos em publicidade: os gastos com publicidade cresceram 16% em 2011 e alcançaram 88,3 bilhões de reais [Barbosa \(2012\)](#). Além disso, ainda em 2011, a *Interactive Advertising Bureau* do Brasil (IAB Brasil) divulgou que as 100 maiores empresas do Brasil investiram 13,4% de seus orçamentos em publicidade no meio digital [Alves \(2012\)](#). Já em 2012, uma pesquisa mostrou que no primeiro semestre, 266 bilhões de dólares foram gastos com anúncios em todo o mundo [Azevedo \(2012\)](#). E até o final do ano, previsões revelam que os gastos com publicidade em dispositivos móveis devem alcançar a marca dos US\$ 6,43 bilhões, valor que deve subir, até 2016, para US\$ 23,6 bilhões [Adnews \(2012\)](#).

Segundo [Gratton \(2002\)](#), *location-based advertising* permite aos usuários com dispositivos móveis receberem conteúdo baseado na posição geográfica dos mesmos. Esse conteúdo pode ser mensagens publicitárias, cupons promocionais ou até mesmo ofertas, que podem ter como origem estabelecimentos próximos ao usuário. Este tipo de publicidade, que muda baseada na localização, cria a ideia de que um anunciante poderia “atingir” um

cliente no momento que ele está mais apto a comprar, sob o aspecto da proximidade física. Para Shannon et al. (2009), um anúncio personalizado e bem projetado pode ser mais persuasivo para um cliente do que um *outdoor* tradicional. Por exemplo, usuários que estejam em um shopping ou em uma área urbana comercial poderiam ser avisados de ofertas locais que potencialmente os interessaria. Outro exemplo, encontrado em uma matéria do Chief Marketer¹, descreve um serviço gratuito criado para enviar mensagens de texto que oferecem conteúdo relacionado aos filmes que estão em cartaz, como trailers e toques de celular. Para isto, basta que os usuários estejam a certa distância das salas de cinema, algo em torno de dez metros. A ideia é abordar os usuários através de meios alternativos, indo além da TV, jornal e Web.

Por um lado, LBA normalmente é realizado em *broadcast*, sem identificação das pessoas – a única informação disponível é a própria localização. Por outro, as abordagens que buscam identificar o perfil são baseadas no fornecimento explícito de informações pelo próprio usuário. Porém, tais abordagens podem ser inviabilizadas em um contexto de recomendação em massa, quando muitas vezes o usuário torna-se impessoal, dada a massificação do seu propósito. Como resultado, pode-se ter um anúncio intrusivo e um cliente bastante insatisfeito, pois o perfil buscado não era compatível com o tal usuário. Deve-se, portanto buscar estratégias que justifiquem os altos investimentos mencionados anteriormente.

Nesse sentido, faz-se necessário ter estratégias de inferência de perfil a partir de dados obtidos automaticamente, como por exemplo, via monitoramento de dispositivos móveis. Porém, as soluções que utilizam apenas dados de monitoramento não apresentam resultados satisfatórios, em especial quando considerado o desvio padrão com base na amostra, ou seja, existem subperfis de usuários com um mesmo padrão de deslocamento. Sendo assim, é necessário melhorar a qualidade do perfil do usuário, combinando a utilização de dados relativos à sua rotina de deslocamento com dados públicos relativos à sua localização. Os dados de deslocamento podem ser obtidos, por exemplo, através do já citado monitoramento de seus dispositivos, ou até mesmo através da utilização de serviços sensíveis ao contexto da localização. Em relação aos dados públicos de localização, um exemplo relevante, dito anteriormente, é a existência de dados georreferenciados, como é o caso dos setores censitários, fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

Todo este esforço deve-se ao fato de que quando se busca anunciar, pretende-se obter o melhor resultado, já que anualmente são feitos investimentos na cifra dos bilhões. Portanto, acredita-se que a adição de um mecanismo mais sofisticado, que realize uma disseminação mais personalizada tenha o potencial de aumentar o retorno sobre o investimento (ROI) em anúncios e compatibilizar, de um modo mais eficaz, o perfil buscado com o perfil encontrado.

¹<http://goo.gl/jYDpd>

Como exemplo, podemos citar [Walczak \(2005\)](#), que descreve um produto que, através de personalização e de uma abordagem *opt-in* (formato não intrusivo de publicidade), tem o potencial de dar uma boa contribuição à publicidade móvel. Desta forma, pretende-se desenvolver uma abordagem para inferência de perfis de usuários, gerados por dados georreferenciados e por um histórico de localizações, e recomendar anúncios que sejam compatíveis a estes perfis. De forma geral, seu objetivo é melhorar a eficácia dos anúncios exibidos em dispositivos móveis, dado que se busca ter um perfil mais próximo do perfil real do usuário.

No levantamento bibliográfico realizado, foram identificados poucos trabalhos neste sentido. Desta forma, pretende-se desenvolver uma proposta que culmine em um serviço inovador, que seja útil e menos intrusivo para a população e que também esteja relacionado à Inteligência Artificial, mais especificamente com criação de perfis de modo automatizado e com sistemas de recomendação. A Tabela 1.1 sumariza as principais características almejadas.

Tabela 1.1: Principais características do trabalho proposto.

Anonimato dos usuários
Gerar perfil automaticamente, sem a interferência do usuário
Utilização de dados georreferenciados
Histórico de localização
Flexibilidade na obtenção dos dados de localização

Fonte: Autor, 2012.

Devido ao fato de muitos dispositivos móveis, no Brasil, ainda não possuírem GPS embutido, uma limitação em nossa abordagem é que para criar o histórico de localização de um usuário, outras estratégias devem ser utilizadas, como uma técnica chamada Triangulação. Outro problema que deve ser destacado é o fato de abordar corretamente os usuários com a publicidade, pois, assim como em *e-mails*, os usuários podem tratar as mensagens recebidas como *spams*. Vale destacar que em nosso trabalho não foram tratados aspectos de privacidade com a utilização do serviço, ficando isto a cargo de trabalhos futuros.

1.1 Contextualização

Para se ter uma ideia mais clara da aplicabilidade do que está sendo proposto, podemos analisar a situação em que, diariamente, milhares de pessoas dependem do serviço de transporte público. À mesma proporção que dependem deste tipo de serviço, elas também são expostas a vários tipos de mídias exteriores, como letreiros luminosos nos pontos de ônibus, *backbus* ou *outbus* (publicidade presente na parte traseira dos ônibus) e também em televisões dentro dos mesmos, que informam sobre as notícias mais importantes mas que também veiculam propaganda. A existência de um sistema de recomendação de anúncios que, através de um histórico de localização dos usuários e de informações georreferenciadas da região de

atuação, possibilitaria recomendar conteúdo menos intrusivo e com uma série de pontos positivos:

- Anúncios mais personalizáveis;
- Maior potencial de retorno de investimento (ROI) em anúncios;
- Identificação automática e mais precisa do perfil dos usuários anônimos.

1.2 Objetivos

1.2.1 Objetivo Principal

O objetivo geral deste trabalho é criar e refinar perfis de pessoas com base em um conjunto de dados históricos de localização e em dados georreferenciados, resultando numa proposta que contribua na melhoria da eficácia de anúncios sensíveis à localização e proporcione um serviço de envio de conteúdo mais personalizado.

1.2.2 Objetivos Específicos

Para atingir o objetivo principal pretendido, serão utilizadas técnicas de Inteligência Artificial relacionadas a sistemas de recomendação automática e criação de perfis. Mais precisamente, pretendemos chegar ao objetivo principal através dos seguintes objetivos específicos:

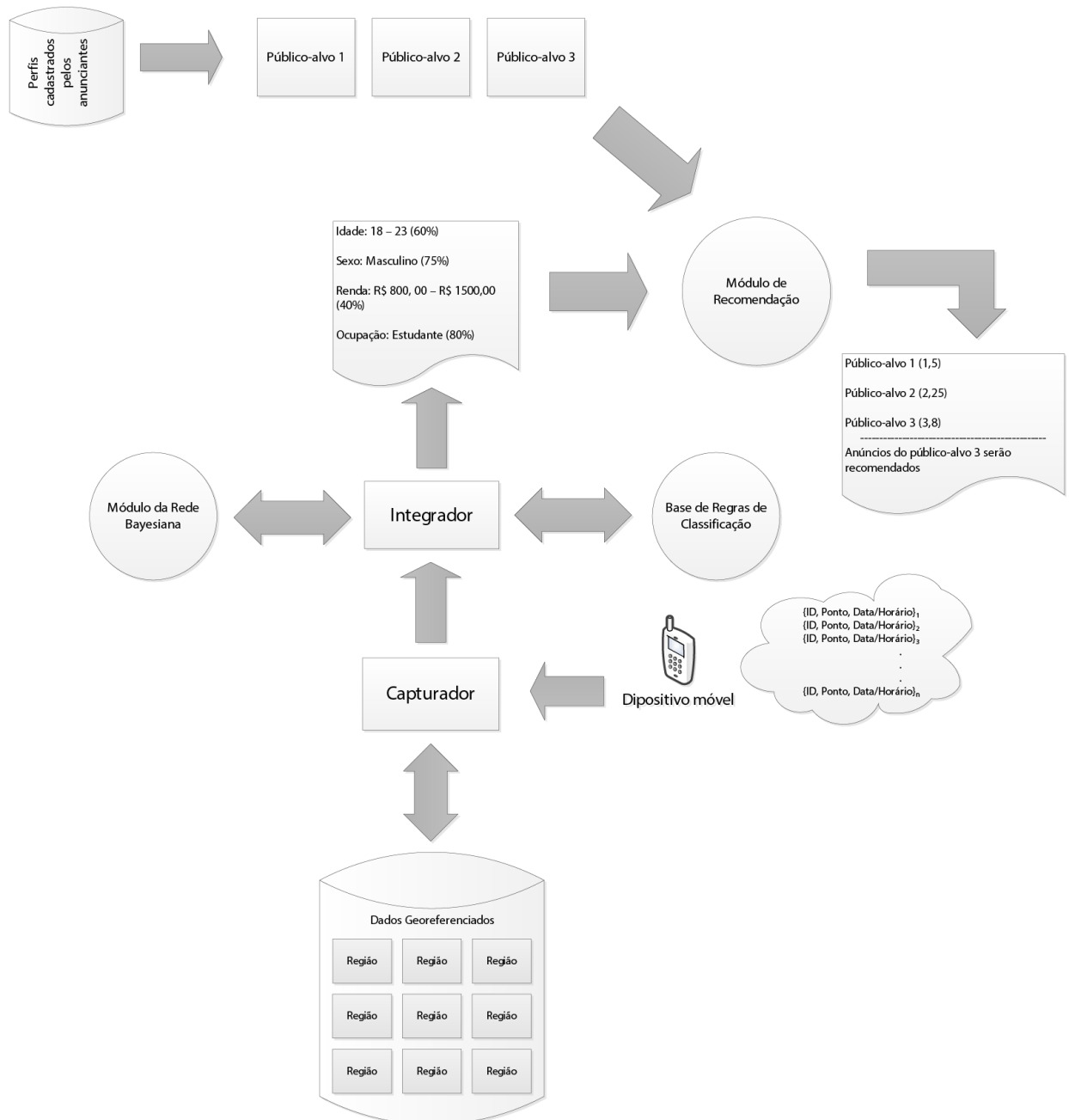
- Estudar o estado da arte atual relativo a: (i) sistemas de anúncio baseado em localização; (ii) tecnologias envolvidas na implementação desses sistemas; e (iii) algoritmos e técnicas existentes para criação automática de perfis;
- Desenvolver um algoritmo para criação de perfil dos usuários extraindo informações a partir de dados georreferenciados em conjunto com históricos de localização;
- Implementar um serviço de recomendação de conteúdo para usuários de dispositivos móveis baseado em sua localização e em dados históricos;
- Realizar validação do serviço em cenários reais.

INTRODUÇÃO

1.3 Visão geral da proposta

O esquema geral do nosso trabalho pode ser visualizado na Figura 1.1. Posteriormente, no Capítulo 4, teremos esta figura explicada detalhadamente em partes menores.

Figura 1.1: Diagrama que apresenta a ideia geral da proposta do trabalho.



Fonte: Autor, 2012.

1.4 Estrutura da dissertação

O restante dessa dissertação obedece à seguinte estrutura:

- No Capítulo 2 é apresentada a fundamentação teórica inerente a esse trabalho;
- No Capítulo 3 são abordados os trabalhos relacionados obtidos no levantamento do estado da arte;
- No Capítulo 4 é mostrada a ideia central do trabalho, apresentando o modelo desenvolvido.
- No Capítulo 5 apresenta-se uma ideia geral do método *Goal/Question/Metric* e mostram-se alguns aspectos do estudo de caso;
- No Capítulo 6, temos a descrição do estudo de caso realizado para validar o trabalho.
- No Capítulo 7, temos, por fim, o fechamento do trabalho, com a devidas conclusões e perspectivas de trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo é dada uma visão geral sobre as principais áreas do âmbito deste trabalho. Dessa forma, são apresentados conceitos sobre *Location-based Services*, com foco em *Location-based Advertising*, assim como Redes Bayesianas, Árvores de Decisão, e ainda, uma pequena descrição sobre Métricas de Similaridade.

2.1 Location-based Services

Segundo Schiller & Voisard (2004), o termo *location-based services* (LBS's) representa serviços que integram a localização ou posição de um dispositivo móvel com outras informações para prover valor agregado ao usuário ou como um conceito que denota aplicações que integram localização geográfica (e.g. coordenadas espaciais) com serviços. Exemplos de tais aplicações incluem serviços de emergência, sistemas de navegação em carros, planejamento de passeios turísticos e “mapas amarelos” de informação (combinação de páginas amarelas com mapas). Com o desenvolvimento da comunicação móvel, estas aplicações representam um novo desafio tanto conceitual quanto técnico. Ainda segundo Schiller & Voisard (2004), muitas destas aplicações serão parte da rotina das pessoas, através de computadores, telefones celulares entre outros. Já para Küpper (2005), não existe nem uma definição comum nem uma terminologia comum para o termo *location-based services* ou serviços baseados em localização (tradução livre) na área de comunicação móvel. Os termos *location-based service*, *location-aware service*, *location-related service* e *location service* são usados indistintamente. Uma razão para esse dilema poderia residir no fato de que a nomenclatura de tais serviços é determinada por diferentes comunidades, especialmente o setor de telecomunicações e a área de computação ubíqua. Ainda segundo Küpper (2005), a *GSM Association*, um consórcio de 600 operadoras da rede GSM, simplesmente define LBS's como serviços que usam a localização de um alvo para adicionar valor aos serviços. Ainda são apresentados três exemplos: selecionar pontos próximos de interesse, mostrar a localização de um alvo no mapa e ativar automaticamente um serviço quando um alvo entra ou sai de uma localização pré-definida.

Em Schiller & Voisard (2004) é descrito que, historicamente em sistemas de posicionamento tradicional, a informação de localização surgiu dos sistemas de satélite. Contudo, o amplo interesse em serviços baseados em localização só começou a ser impulsionada no final dos anos 90, quando um novo tipo de tecnologia de localização e um novo nicho de mercado em serviços de dados foram desencadeados pelas operadoras de rede móvel. Aproximadamente em 1997, as redes móveis foram largamente implantadas na Europa, Ásia e Estados Unidos e os lucros dos serviços de telefonia provaram ser significativos para as operadoras móveis. No entanto, embora os serviços de voz móveis continuassem a ser o maior gerador de lucros, o crescimento da telefonia móvel era limitada,

e o preço por minuto estava sendo barateado. Conseqüentemente, as operadoras tiveram que começar a procurar por meios alternativos de estabilizar suas finanças, ou seja, áreas com potencial de crescimento futuro. A solução para captar benefícios financeiros adicionais, além dos serviços de voz, foi oferecer serviços de dados, muitos dos quais aprimorados pela localização. Além disso, para melhorar os serviços de dados, as operadoras passaram a investir em novas tecnologias e a procurar meios de otimizar a experiência do usuário neste novo domínio.

Dessa forma, a localização do usuário passou a ser uma importante variável a ser considerada: não só permite às companhias desenvolverem serviços completamente novos como também tem o potencial de fazer muitas mensagens e serviços mais relevantes aos clientes, ajustados ao contexto como por exemplo, previsão do tempo para a região do usuário. Este tipo de informação passou a ser tão relevante que as operadoras estão começando a considerá-la como um “terceiro ativo”, juntamente com voz e transmissão de dados [Schiller & Voisard \(2004\)](#).

2.1.1 Location-based Advertising

Dentre os vários tipos de serviços baseados em localização, como descrito na Tabela 2.1, um que pode ser destacado é *Location-based Advertising* (LBA) ou Anúncio baseado em Localização. De acordo com [Kölmel & Spiros \(2002\)](#), o termo LBA pode simplesmente ser entendido como uma publicidade que muda baseada na localização do usuário, e que tem sido o viés da internet *wireless* que mais tem se falado. A ideia é que um anunciante tenha contato com cliente no momento em que ele está mais propício a comprar, enviando, por exemplo, mensagens a usuários que estejam passeando em um shopping ou passando por uma área de comércio, informando-os das ofertas locais. [Dao et al. \(2012\)](#) afirma que LBA é uma nova forma de publicidade que vem despertando atenção, em que os provedores de serviços rastreiam a localização dos usuários-alvo através de seus dispositivos móveis, e os provê de anúncios específicos à localização.

Para [Bruner & Kumar \(2007\)](#), estamos em um momento em que podemos entrar em contato com as pessoas a qualquer hora e em qualquer lugar, devido à ubiquidade proporcionada pelos telefones celulares, principalmente aqueles com *Global Positioning System* (GPS). A tecnologia atual detém o potencial que permite aos comerciantes enviar informações e ofertas aos consumidores com base na sua proximidade a locais onde os produtos deles estão disponíveis. Dessa forma, [Bruner & Kumar \(2007\)](#) descreve LBA, de forma geral, como algo que se refere à informação controlada por comerciantes/anunciantes, especialmente adaptada para o lugar onde os usuários acessam um meio de publicidade. Definindo dessa forma, vemos que LBA não é um conceito novo, já que o conteúdo dos *outdoors* é específico do lugar onde eles estão localizados. Por outro lado, há um meio completamente novo que permite aos usuários serem expostos a publicidade em seus dispositivos móveis. Desta forma, LBA faz parte do chamado

real-time marketing, encontrando as necessidades do cliente no momento e no lugar adequados.

Tabela 2.1: Classificação dos serviços baseados em localização.

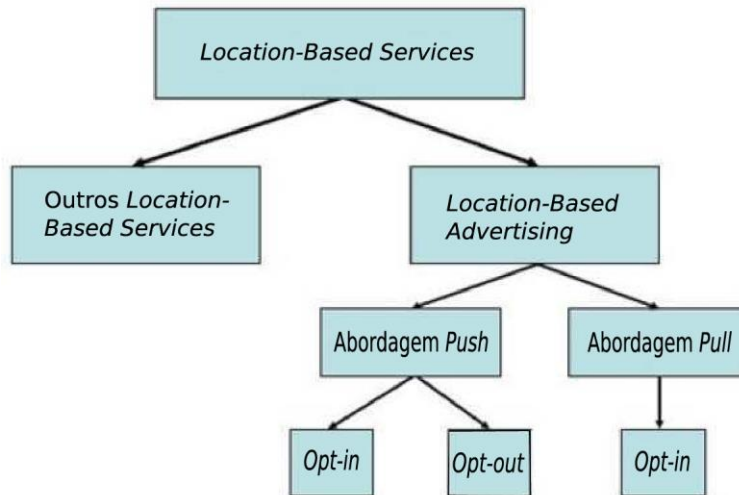
Categorias de LBS	Tipos de LBS padronizados
Serviços de segurança pública	Serviços de emergência Serviços de alerta de emergência
Serviços de rastreamento	Rastreamento pessoal Gestão de frotas Gestão de bens
Monitoramento do tráfego	Relatórios de congestionamentos do tráfego
Roteamento de chamada aprimorado	Assistência na estrada Roteamento para empreendimento comercial mais próximo
Serviços de informação baseado na localização	Informação do transporte público e do tráfego Visita aos pontos turísticos de uma cidade Anúncio baseado na localização Páginas amarelas em dispositivos móveis Informações sobre o tempo
Serviços comunitários e de entretenimento	Jogos Encontrar um amigo Encontros de casais Bate-papo Busca de rota Onde estou?

Fonte: Adaptado de Schiller & Voisard, 2004.

Push e Pull

Como ilustrado na Figura 2.1, temos duas abordagens possíveis em LBA: a abordagem *push* e a abordagem *pull*. Para Bruner & Kumar (2007), apesar desses conceitos serem bem conhecidos no marketing tradicional, eles são ligeiramente diferentes em LBA, se restringindo ao contexto de anúncios e ofertas.

Na abordagem *push*, o anunciante trabalha em conjunto com as operadoras de rede para enviar anúncios aos usuários, a partir da localização do mesmo. Dentro dessa abordagem, há mais duas divisões, que se tornaram populares nos últimos anos: *opt-out* e *opt-in*. *Opt-out* sugere que os anunciantes enviem anúncios a qualquer um que eles desejem até o momento em que cada usuário informe explicitamente que não deseja mais recebê-los. Já o *opt-in*, envolve a autorização prévia do usuário para que as mensagens sejam enviadas a eles. Apesar de ser considerado intrusivo este tipo de abordagem alcança um público maior, na maioria das vezes, de forma indiscriminada. Esta tem sido a preferida dos anunciantes, devido ao seu alcance e simplicidade, apesar dela ser amplamente considerada como uma nova forma de *spam*.

Figura 2.1: Tipos de *Location-based Advertising*.

Fonte: Adaptado de Bruner & Kumar, 2007.

Mas segundo Chuang & Krabuanrat (2007), muitas vezes decisões de compra são feitas por impulso e *location-based push advertising* é particularmente adequado para este propósito. De fato, se apropriadamente implementado, anúncios do tipo *push* podem ser benéficos para os comerciantes, como seria por exemplo, se alguém recebesse um cupom, no formato de mensagem, concedendo 60% de desconto em uma loja de roupas importadas, o que dificilmente seria tratado como um incômodo. Em suma, os anúncios precisam ser suficientemente atraentes e relevantes para usufruírem dessa abordagem. Para exemplificar, vamos imaginar que alguns espectadores estejam esperando a próxima sessão em um cinema, e recebem mensagens de texto informando-os que eles podem baixar gratuitamente conteúdo relacionado aos filmes que estão em cartaz. Neste caso, para a abordagem *opt-in*, os expectadores deveriam ter indicado previamente o interesse deles em receber essas mensagens do cinema, enquanto que para *opt-out* os usuários não necessariamente precisariam ter feito isso.

A outra abordagem, chamada *pull*, ocorre quando os usuários solicitam alguma informação ou usam algum serviço, e são expostos a mensagens comerciais. Nestas situações, os usuários estão buscando uma informação em vez da informação estar em busca deles. De acordo com Chuang & Krabuanrat (2007), esta abordagem vem sendo considerada a mais adequada para LBA, pois além de conseguir capturar a localização do usuário quando ele está buscando informação, o problema de *spam* é totalmente eliminado. Além disso, não é solicitada uma pré-identificação do usuário ou qualquer outra interação com o anunciante. Ou seja, é inteiramente possível para um anunciante ter um novo cliente usando LBA contanto que o mesmo esteja dentro da sua vizinhança ou buscando informação.

Como exemplo, vamos imaginar um turista chegando a uma cidade. Usando seu dispositivo móvel, ele pode usar algum serviço que mostre os restaurantes mais próximos a sua localização. No processo de utilização do serviço, pode ser apresentada uma oferta que dá 50% de desconto para os próximos 10 clientes que chegarem a um determinado restaurante japonês. A Tabela 2.2 apresenta os principais aspectos das duas abordagens citadas.

Tabela 2.2: Push vs Pull

	Pull	Push
Vantagens	Menos invasivo Marketing mais direcionado Menor custo para os anunciantes Capacidade para reconhecer os perfis dos clientes	Alcance maior de público
Desvantagens	Alcance menor de público	Problemas de <i>spam</i> Altos custos
Fatores de sucesso	Capacidade de adquirir conhecimento através do perfil dos clientes	Anúncios atraentes o bastante para não afastar os clientes

Fonte: Autor, 2012

2.2 Redes Bayesianas

De acordo com [Russell & Norvig \(1995\)](#) e [Heckerman \(1996\)](#), uma rede bayesiana é um grafo orientado em que cada nó é identificado com informações de probabilidade quantitativa. Em outras palavras, podemos definir do modo a seguir:

- Um conjunto de variáveis aleatórias formam os nós da rede, em que estas variáveis podem ser discretas ou contínuas.
- Um conjunto de vínculos (setas) orientados conecta pares de nós. Uma seta do nó X até o nó Y implica em dizer que X é pai de Y ou X tem uma influência direta em Y.
- Cada nó X_i tem uma distribuição de probabilidade condicional definida por $P(X_i | Pais(X_i))$, que quantifica o efeito dos pais sobre o nó.
- O grafo não tem nenhum ciclo orientado.

Segundo [Russell & Norvig \(1995\)](#), o conjunto de nós e arestas, que formam a rede, especifica os relacionamentos de independência condicional que são válidos no domínio. E ainda descreve:

Uma vez que a topologia da rede bayesiana está definida, só precisamos especificar uma distribuição de probabilidade condicional para cada variável, dados seus pais. A combinação da topologia com as distribuições condicionais é suficiente para especificar (de forma implícita) a distribuição conjunta total para todas as variáveis.

Para [Darwiche \(2009\)](#) a rede bayesiana, como uma ferramenta de representação, é bastante atraente por três motivos:

- É uma representação consistente e completa, garantindo definir uma distribuição de probabilidade única sobre as variáveis da rede.
- A rede bayesiana é modular, no sentido de que a sua consistência e integridade são garantidas por meio de testes localizados que se aplicam apenas às variáveis e suas causas diretas.
- Por último, ela é uma representação compacta, uma vez que permite especificar uma distribuição de probabilidade de tamanho exponencial utilizando um número polinomial de probabilidades.

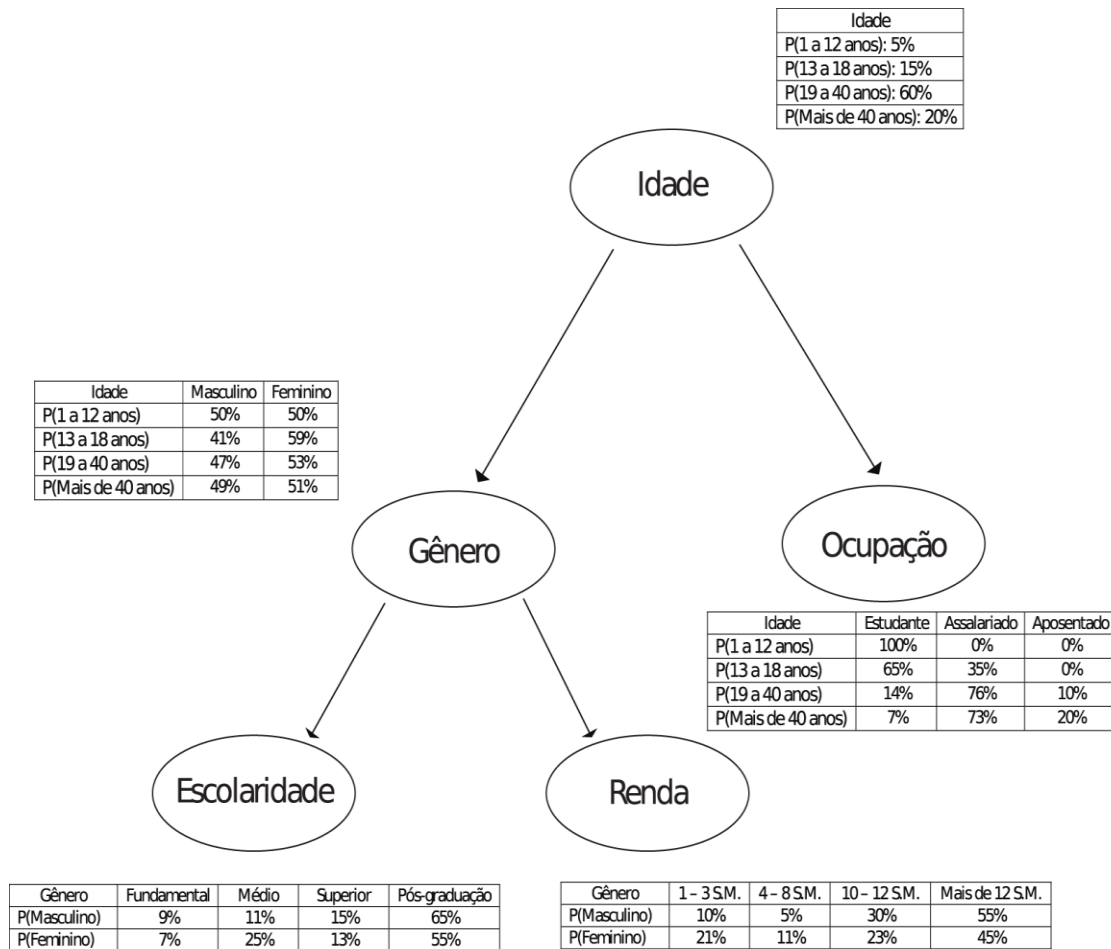
Para ilustrar o que foi descrito, será apresentado um exemplo a seguir. Ele ilustra a situação em que desejamos saber o perfil, de modo probabilístico, dos usuários de transporte público de uma cidade, dado um histórico de deslocamento diário e informações do censo sobre esta cidade. O deslocamento nos fornece informações sobre como o usuário se comporta em sua rotina nos ônibus, permitindo inferir, por exemplo, onde ele trabalha e/ou estuda e também onde mora. Já as informações do censo, que poderiam ser generalizadas como informações georreferenciadas, permitem saber qual o perfil da população, como idade, número de filhos, escolaridade, renda, entre outras informações. Dentre os vários aspectos citados, vamos supor que temos interesse somente nos atributos idade, gênero, escolaridade, renda e ocupação. Sabendo que o censo mostra a distribuição de probabilidade para cada um dos valores assumidos pelos atributos e que idade determina ocupação e gênero, e este por sua vez determina renda e escolaridade, uma rede bayesiana para a situação descrita pode ser vista na [Figura 2.2](#).

Esse exemplo nos mostra que, fixada uma pessoa, ela tem 15% de probabilidade de ter entre 13 e 18 anos. Já no atributo gênero, dado que ela está na faixa de 13 a 18 anos, ela tem 41% de chance de ser do sexo masculino e 59% de ser do sexo feminino. Já na escolaridade, se ela for do sexo masculino, ela tem 11% de chance de ter o ensino médio e 15% o ensino superior. A mesma lógica se aplica aos demais atributos.

No tocante à realização de inferência em redes bayesianas, é necessário calcular a distribuição de probabilidade posterior para um conjunto de variáveis de consulta, dado que algum evento foi observado. Voltando ao exemplo da rede bayesiana do perfil probabilístico,

poderíamos observar o evento em que temos 100% de certeza que o morador é assalariado, dado o universo assalariado e estudante. Inserindo esta evidência, poderíamos investigar como isso é refletido nos demais atributos, como a idade ou o gênero.

Figura 2.2: Típica rede bayesiana, que mostra a relação probabilística entre os atributos do perfil de um usuário a partir de dados do censo.



Fonte: Autor, 2012.

A ideia de utilizar redes bayesianas no presente trabalho pode ser justificada tanto pela representação consistente e compacta, quanto por esta flexibilidade de inserir evidências e ver como os demais atributos da rede se comportam. Em nosso caso, utilizamos redes bayesianas para representar e guardar os perfis probabilísticos dos usuários.

2.3 Árvores de Decisão

Segundo Nilsson (1998), uma árvore de decisão pode ser definida, de forma geral, como uma árvore cujos nós internos são testes e cujos nós folhas são categorias. Um conceito equivalente é encontrado em Russell & Norvig (1995), que afirma que as árvores de decisão tomam como entrada um objeto ou situação descritos por um conjunto de atributos e retorna uma decisão. Tanto os atributos de entrada quanto os valores de saída podem ser discretos ou contínuos.

Retomando Russell & Norvig (1995), ele afirma:

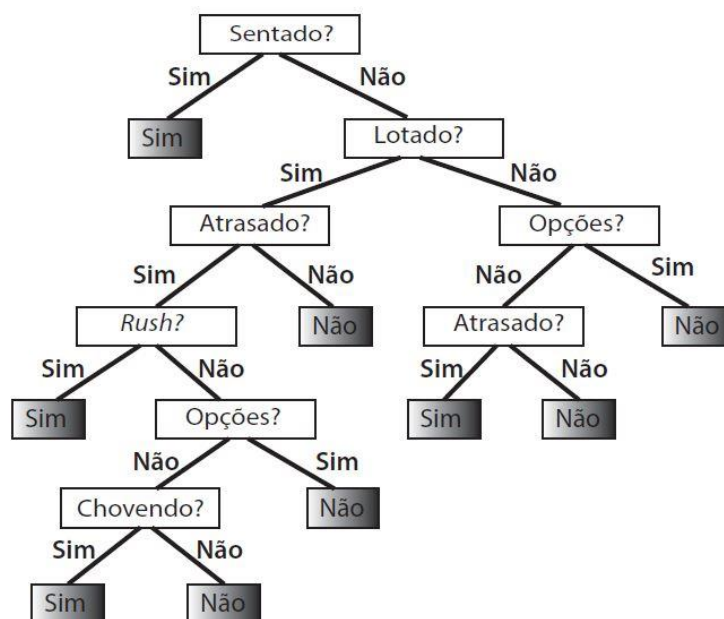
Uma árvore de decisão alcança sua decisão executando uma sequência de testes. Cada nó interno corresponde a um teste do valor de uma das propriedades, e as ramificações a partir do nó são identificadas com os valores possíveis de teste. Cada nó de folha na árvore especifica o valor a ser retornado se aquela folha for alcançada.

O exemplo a seguir descreve uma situação em que uma pessoa decide se deve ou não pegar o ônibus que está parado no ponto. Inicialmente, vamos descrever cada um dos atributos que fazem parte da figura:

1. *Opções*: Se há outras opções de ônibus disponíveis para chegar ao destino.
2. *Atrasado*: Se a pessoa está atrasada para algum compromisso.
3. *Lotado*: Se a quantidade de pessoas no ônibus está acima do que seria permitido.
4. *Chovendo*: Se está chovendo.
5. *Sentado*: Se o ônibus dispõe de lugares para a pessoa sentar.
6. *Rush*: Se o horário é aquele em que os congestionamentos são comuns, normalmente de 7h às 9h e de 17h às 20h.

A árvore de decisão para esse domínio é ilustrada na Figura 2.3. Note que a árvore não utiliza os atributos *EstadoDoÔnibus* ou *ValorTarifa*, considerando-os irrelevantes. As entradas são processadas pela árvore a partir da raiz e seguindo a ramificação apropriada até alcançar uma folha. Por exemplo, uma situação com *Sentado = Não*, *Lotado = Sim* e *Atrasado = Não* será classificada como negativo, ou seja, que a pessoa deve aguardar o próximo ônibus. Em nosso trabalho, como poderá ser visto mais adiante, utilizamos árvores de decisão para tentar inferir a ocupação de uma pessoa, dado sua rotina diária. Assim, tentamos descobrir se a pessoa se trata de um assalariado, estudante ou até mesmo ambos.

Figura 2.3: Árvore de decisão que indica se uma pessoa deve pegar um ônibus.



Fonte: Autor, 2012

2.4 Métricas de Similaridade

Métricas ou medidas de similaridade nos auxiliam a medir quão próximos dois elementos estão, sejam estes vetores, documentos ou até mesmo um conjunto de termos e um documento. Para [Egghe & Michel \(2002\)](#), métricas de similaridade são bem conhecidas e usadas na literatura de *Information Retrieval* (IR) ou Recuperação de Informação (RI). Vale citar trabalhos bastante relevantes que tratam destas medidas, como [Boyce et al. \(1995\)](#), [Tague-Sutcliffe \(1995\)](#), [Grossman & Frieder \(1998\)](#), [Losee \(1998\)](#) e os livros clássicos de [Salton & McGill \(1987\)](#) e [Van Rijsbergen \(1979\)](#).

De forma geral, se temos um conjunto-universo Ω do qual subconjuntos A, B, \dots são gerados, então para qualquer $A, B \subset \Omega$, podemos definir uma métrica de similaridade D (um número positivo $D(A,B)$), que expressa o grau de similaridade entre dois conjuntos, ou seja:

$$D : \Omega \times \Omega \rightarrow \mathbb{R}^+ \quad (2.1)$$

$$(A, B) \rightarrow D(A, B)$$

Para a relação acima, alguns requisitos precisam ser obedecidos: $D(A,B)$ deve ser mínimo (0 na maioria dos casos) se $A \cap B = \emptyset$ e deve ser máximo (1 na maioria dos casos) se $A = B \neq \emptyset$, estabelecendo uma variação, em grande parte dos casos, no intervalo $[0, 1]$. Se D representa o grau de similaridade, podemos dizer que $1 - D$ representa uma medida de dissimilaridade. Para aferir o grau de similaridade, dispomos de vários tipos de coeficientes, como Jaccard, Dice e Cosseno de Salton. Um estudo que estabelece relações entre as métricas citadas é realizado em [Egghe \(2009\)](#). Em [Egghe & Michel \(2002\)](#), é explicado como estas mesmas métricas podem ser utilizadas em termos de similaridade de vetores.

2.4.1 Similaridade por cosseno

Para calcular a similaridade entre dois vetores $\vec{V}(d_1)$ e $\vec{V}(d_2)$, usa-se, de forma geral, uma medida baseada no cálculo do cosseno, também conhecida por cosseno de Salton, definida por:

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)| |\vec{V}(d_2)|} \quad (2.2)$$

no qual o numerador representa o **produto escalar**, também chamado de produto interno, dos vetores $\vec{V}(d_1)$ e $\vec{V}(d_2)$, e o denominador é o produto das **distâncias Euclidianas Manning et al. (2008)**. O produto interno \vec{x} e \vec{y} de dois vetores, de M componentes, é definido como

$\sum_{i=1}^M x_i y_i$. A distância Euclidiana é definida por $\sqrt{\sum_{i=1}^N \vec{V}_i^2}$ de um vetor \vec{V} de N

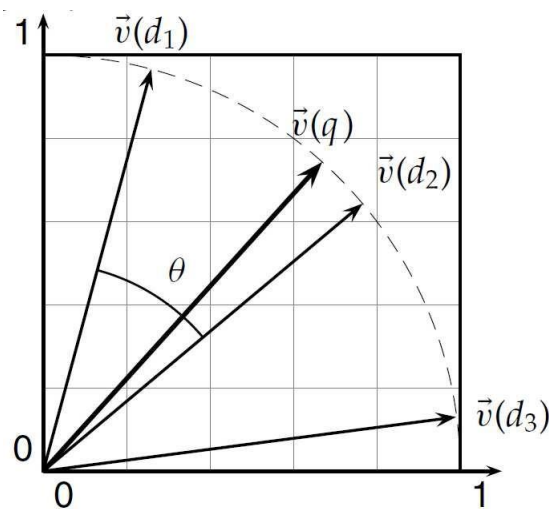
componentes. Podemos reescrever a Equação 2.2 em termos de vetores unitários, em que

$\vec{v}(d_1) = \frac{\vec{V}(d_1)}{|\vec{V}(d_1)|}$ e $\vec{v}(d_2) = \frac{\vec{V}(d_2)}{|\vec{V}(d_2)|}$, obtendo:

$$\text{sim}(d_1, d_2) = \vec{v}(d_1) \cdot \vec{v}(d_2) \quad (2.3)$$

Dessa forma, a Equação 2.3 pode ser vista como o produto interno de versões normalizadas dos dois vetores. Esta medida é o cosseno do ângulo θ entre os dois vetores, ilustrado na Figura 2.4.

Figura 2.4: A similaridade por cosseno: $\text{sim}(d_1, d_2) = \cos\theta$.



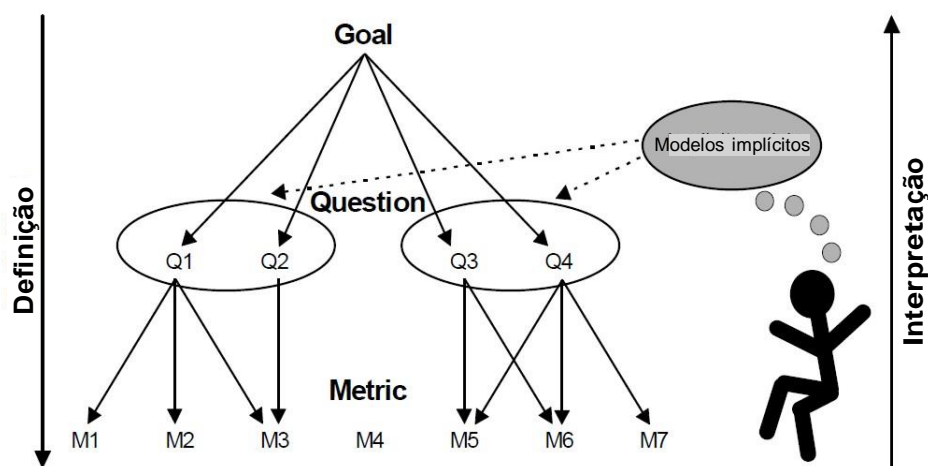
Fonte: Adaptado de Manning et al., 2008.

Um exemplo, no contexto do trabalho, poderia ser a situação em que desejamos recomendar os anúncios de uma loja de roupas para um usuário de ônibus, resgatando aquele exemplo da seção 2.2. Dado que a loja informou qual o público-alvo (para este exemplo iremos considerar somente idade, gênero e renda), poderíamos medir a similaridade entre este e o perfil do usuário (dado que estamos de posse do mesmo). Para isso, basta criarmos um vetor de N posições, em que N é definido pela quantidade de valores que o atributo pode assumir, para cada um dos atributos, somar as similaridades, e ver quão próximo de 3 chegamos, dado que a similaridade de cada atributo pode variar de 0 a 1. Estabelecido um limiar de interesse para a similaridade, vemos se é viável ou não enviar anúncios/ofertas daquela loja para aquele usuário. No presente trabalho, utilizaremos a similaridade para verificar justamente quão próximos estão os dois perfis mencionados.

2.5 O método Goal/Question/Metric

Goal/Question/Metric (GQM) representa uma abordagem sistemática para adequar e integrar objetivos aos modelos do processo de software, produtos e perspectivas de qualidade, baseado em necessidades específicas do projeto e da organização [Basili et al. \(1994\)](#). O princípio por trás deste método é que a avaliação seja orientada a objetivos e estes objetivos sejam transformados em atividades que possam ser medidas durante a execução de um projeto. Ou seja, o método supracitado define um certo objetivo ou conjunto de objetivos, refina-o(s) em perguntas e define métricas para responder estas perguntas. Respondendo as questões, os dados avaliados podem ser analisados para identificar se os objetivos foram atingidos [Van Solingen & Berghout \(1999\)](#). Dessa forma, o GQM define métricas através de uma perspectiva *top-down* mas analisa e interpreta a avaliação dos dados de modo *bottom-up*, ilustrado na Figura 2.5.

Figura 2.5: O método GQM.



Fonte: Adaptado de [Basili & Weiss, 1984](#).

Em um guia prático do GQM, proposto por [Van Solingen & Berghout \(1999\)](#), temos:

O modelo GQM inicia *top-down* com a definição de um objetivo de avaliação explícito. Este objetivo é refinado em várias questões que dividem o problema em seus principais componentes. Cada questão é, então, aprimorada em métricas que devem prover informação para responder cada uma das questões. Como as métricas foram definidas com um objetivo explícito em mente, a informação provida pelas métricas deve ser interpretada e analisada com relação ao objetivo, para concluir se foi ou não alcançado.

Para facilitar a compreensão, vamos imaginar a situação em que desejamos criar uma função, em uma linguagem de programação qualquer, que nos informe a quantidade de alunos aprovados em História, em uma dada turma, antes de eles terem ido para a reavaliação. Neste caso, nosso **Goal** pode ser definido por: “Medir a quantidade de alunos que obtiveram médias acima de 7,0, antes da reavaliação, para a disciplina de História”. A partir desse objetivo, podemos fazer a seguinte pergunta: “Inicialmente, qual a proporção de alunos aprovados em História?”. Por fim, definimos a seguinte métrica: $\frac{\text{quantidade de alunos com média acima de 7,0}}{\text{quantidade de alunos na turma}}$. Para este exemplo, temos a impressão de que se trata de algo bem simplório, mas em projetos de desenvolvimento que envolvem várias equipes, isto ajuda a avaliar se as coisas estão sendo feitas corretamente.

Para maiores detalhes sobre o método GQM e suas implicações no desenvolvimento e qualidade de software, consultar [Basili et al. \(1994\)](#), [Van Solingen & Berghout \(1999\)](#) e [van Solingen et al. \(2002\)](#). Em nosso trabalho, utilizamos este método para estruturar o estudo de caso realizado (ver Capítulo 5).

3 Trabalhos Relacionados

Neste capítulo são apresentados alguns trabalhos relacionados ao tema desta pesquisa. Por se tratar de uma área com muitas aplicações, várias abordagens serão discutidas. Em seguida, é realizada uma breve discussão entre os trabalhos apresentados e essa pesquisa.

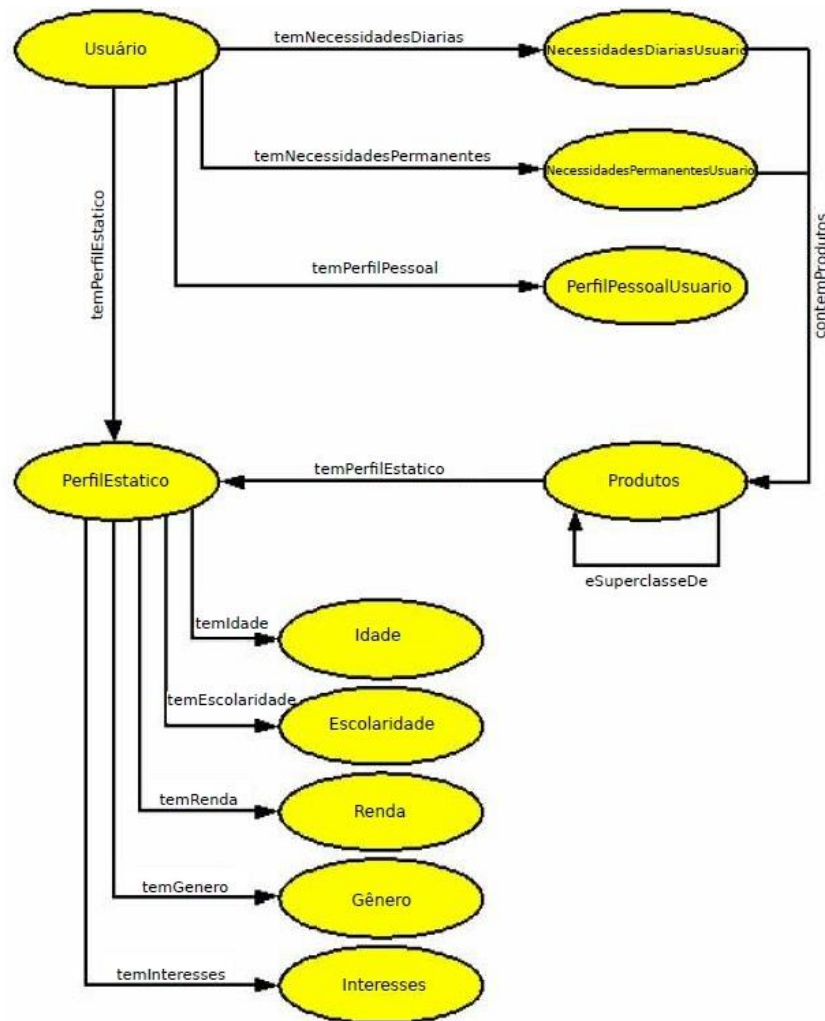
3.1 Estado da Arte

Vários trabalhos começaram a ser desenvolvidos, na tentativa de utilizar a localização como uma nova oportunidade de negócios para alcançar os usuários de modo mais eficaz. Além da localização, outros trabalhos passaram a utilizar diversos mecanismos de personalização na tentativa de adicionar mais valor agregado ao conteúdo anunciado. De acordo com [Walczak \(2005\)](#), simples atributos como idade e gênero poderiam ser analisados ao enviar anúncios.

Buscando realizar um trabalho relacionado à publicidade sensível ao contexto e também em serviços personalizáveis que possam oferecer, aos usuários equipados com dispositivos móveis, uma informação específica, [Liapis et al. \(2008\)](#) propõe uma abordagem, em formato de serviço, chamada Location Aware Shopping Advertisement (LASA). Eles apresentam um serviço assinável, de baixo custo, personalizável, não-intrusivo e baseado em redes *Wi-Fi*, que envia publicidade a dispositivos que estejam essencialmente conectado a uma *Infostation*¹. Se o dispositivo móvel do usuário permitir, o serviço também faz uso da tecnologia *GPS*, o que tornará possível localizá-lo além dos pontos de acesso e mapear outros locais, como sua residência, local de trabalho e casas de praia. Caso não seja possível, o usuário estará, por padrão, localizado naquela área de cobertura à qual ele está conectado. Visando a precisão da representação, esse serviço foi formulado através de ontologias, que são utilizadas para modelar os usuários e os produtos/serviços que serão recomendados (Ver Figura 3.1), assim como as técnicas de raciocínio que são aplicadas para limitar a quantidade de anúncios enviados ao cliente, enviando o que realmente o interessa.

¹Uma *Infostation* é caracterizada por um ou mais pontos de acesso *Wi-Fi*, que cobrem uma certa área, e um servidor que armazena e manipula publicidade e outras informações relacionadas às lojas afiliadas a um *shopping center*.

Figura 3.1: Núcleo da ontologia do LASA.



Fonte: Adaptado de Liapis et al., 2008.

Para a descrição ontológica do usuário, fatores como o posicionamento e a hora daquela posição são associados, permitindo enviar anúncios nos momentos apropriados. Para mapear o perfil, quatro estruturas são utilizadas: *NecessidadesPermanentesUsuario*, *NecessidadesDiariasUsuario*, *PerfilPessoalUsuario* e *PerfilEstatico*.

Para a descrição ontológica dos produtos e serviços é utilizado somente *PerfilEstatico*, que é equiparado ao dos usuários para fazer a recomendação. Todos os produtos e serviços oferecidos ao usuário são representados e exibidos numa estrutura de árvore chamada *p-Tree* (Ver Figura 3.2), que permite uma categorização e subcategorização. É importante mencionar que a categorização não obedece a nenhuma estrutura pré-definida, deixando a cargo do administrador de cada servidor criar uma estrutura que mais se adequa as necessidades de cada anunciante.

Na prática, ao assinar o serviço e rodá-lo no dispositivo, assim que o usuário atingir uma área

de cobertura e se conectar ao servidor, o perfil estático do usuário e do produto/serviço são comparados, e todas as ocorrências compatíveis são enviadas ao usuário.

Com base nos enviados, o usuário pode selecionar os anúncios que deseja visualizar, mas pode também ignorá-los e explicitamente escolher quais deseja, acessando, portanto, outros anúncios que não os recomendados. À medida que o LASA é utilizado, o servidor é atualizado com essas informações, evoluindo o perfil do cliente e os anúncios que futuramente sejam enviados.

Figura 3.2: Classe Produtos seguindo a estrutura *p-Tree*.



Fonte: Adaptado de [Liapis et al., 2008](#).

A utilização de ontologias é uma opção bastante interessante para a representação de conhecimento, o que poderia, com alguns ajustes, ser uma alternativa a rede bayesiana utilizada em nossa abordagem. O ponto em comum é o fato de estabelecer uma relação entre posicionamento e hora para personalizar o conteúdo enviado. Mas, pode-se perceber que um ponto negativo dessa proposta é que ela necessita que as informações que compõem o perfil estático sejam dadas explicitamente pelos usuários, enquanto nossa abordagem cria este perfil sem influência do usuário. Outro ponto que deve ser levantado é a questão da necessidade de um especialista para categorizar os produtos relacionados ao perfil do anunciante.

Uma abordagem bastante semelhante é tratada por Christoforos Panayiotou e George Samaras em [Panayiotou & Samaras \(2006\)](#), no trabalho intitulado “Mobile User Personalization with Dynamic Profiles: Time and Activity”. Eles apresentam um sistema que considera horário e experiência do usuário como atributos fundamentais na personalização de anúncios que são enviados aos dispositivos deles. Utilizar o atributo **hora** significa analisar o horário em que a consulta é realizada (e.g. horário de almoço, café da noite ou da manhã). Já o atributo **vivência** analisa a atividade ou condição a qual o usuário está vivenciando atualmente, ou seja, se o mesmo está em um “dia normal” (dias de trabalho) ou está em um período de férias, situações em que as necessidades mudam completamente. Um bom exemplo seria o sistema identificar, dada uma experiência passada, que o usuário, durante as noites de férias, se interessa por bares ou lugares com *happy hour*. Para criar essa

personalização baseada em tempo e vivência, o primeiro passo seria dividir o dia em várias zonas de horário e armazenar as preferências do usuário para cada zona. Em seguida, seriam atribuídos pesos a cada uma dessas preferências, dada a zona horária. Uma das poucas diferenças entre as abordagens propostas Liapis et al. (2008) e por Panayiotou & Samaras (2006) é que este identifica anomalias na utilização do sistema através dos logs de utilização enquanto aquele mapeia isso para a ontologia.

As variáveis envolvidas na abordagem citada acima são bastante relevantes, pois, visam não somente capturar os dados, mas dar uma interpretação a eles. Por exemplo, ao analisar que o usuário utilizou o sistema às 12:30, isso é inserido em um contexto de “horário de almoço” (zona de horário). Um aspecto equivalente ao atributo **vivência**, descrito acima, pode ser visto na inferência da ocupação, onde nos limitamos a analisar se o dia é um dia de trabalho ou um final de semana. Alguns problemas identificados são: a criação de novas zonas de horário e mudanças de hábito, para o perfil do usuário, é feita a partir da análise de *logs* de utilização do sistema; é necessária a presença de um especialista para determinar um perfil padrão inicial, de acordo com a classe social do usuário; um aspecto que poderia ser avaliado seria a renda do usuário, que não é considerada, visto que no perfil temos somente preferências associadas a zonas de horário.

Buscando na literatura mais algumas abordagens para a personalização de anúncios, uma outra proposta, que tem por objetivo fortalecer a indústria da publicidade direcionada, é apresentada pelos autores de Mazaheri et al. (2010). Eles propõem o *Targeted Adaptive Advertisement* (TAAD), um sistema que alia publicidade com *intelligent Location Based Services* (iLBS). Esse tipo de serviço é baseado em uma política *push* ou *service initiated*, ou seja, provedores de serviço enviam informação sensível à localização aos usuários, considerando interesses pré-definidos e o comportamento deles. Nesse sistema iLBS são usados Redes Bayesianas e o TOP- SIS, um método para tomada de decisões com múltiplos critérios, para, em conjunto, selecionar os anúncios mais adequados aos indivíduos. Em seu funcionamento, o TAAD basicamente utiliza:

- Redes Bayesianas: para a predição das características e comportamentos do usuário. Para cada propaganda que se deseja enviar, é necessário estabelecer inicialmente características e comportamentos associados a clientes daquele tipo de anúncio, criar um formulário que contenha perguntas direcionadas a cada uma dessas características, para ser preenchido pelo usuário, criar uma rede bayesiana para cada característica, em que esta será o nó pai e as respectivas questões e comportamentos os nós filhos. Dessa forma, os clientes que assinam o serviço respondem a esse formulário, e as redes bayesianas são preenchidas. Vale ressaltar que cabe a um especialista estimar cada uma das tabelas de probabilidade.
- O *Fuzzy TOPSIS*: como componente central, gera uma classificação dos anúncios disponíveis de acordo com as características do usuário. A escolha dessa abordagem

extendida do TOPSIS está relacionado ao fato de que os dados utilizados são do tipo *fuzzy*.

Com a combinação de todos esses componentes, gerando essa abordagem personalizável e adaptativa, os autores esperam evitar o envio de publicidade irrelevante para seus clientes. Um trabalho semelhante foi desenvolvido por [Huang & Bian \(2009\)](#). Nele, busca-se recomendar, através da internet, atrações turísticas adequadas a usuários, com base no comportamento de viagens deles. Para isso, além de Redes Bayesianas, é utilizado um método chamado *Analytic Hierarchy Process* (AHP), outro método que auxilia na tomada de decisões complexas.

No trabalho proposto por [Mazaheri et al. \(2010\)](#), a variável **comportamento** é bastante relevante e poderia ser um aspecto a ser adicionado à nossa abordagem. A utilização de redes bayesianas é um ponto em comum para armazenar as características do usuário. Já o método de classificação dos anúncios é diferente, visto que ele utiliza um método mais complexo para tomada de decisões enquanto nós utilizamos uma abordagem mais simples, através da medida de similaridade **Cosseno de Salton**. Um ponto negativo é que essa abordagem utiliza uma metodologia *push*, um método que, de forma geral, não é bem recebida pelos usuários. Além disso, é necessário o preenchimento de um formulário para utilização do serviço, o que num contexto de recomendação em massa pode não ser bem adequado. Mais um vez, temos a necessidade de um especialista, que para esse caso serve para estimar as tabelas de probabilidade e para mensurar quais as características que devem ser avaliadas para um determinado produto. Em nossa abordagem, as tabelas de probabilidade são preenchidas automaticamente, a partir dos dados georreferenciados.

Além dos trabalhos já mencionados, no processo de varredura da literatura, várias outras abordagens foram encontradas. Usando a plataforma de aplicativos do Facebook ², os autores de [Shannon et al. \(2009\)](#) desenvolveram um aplicativo que coleta informações do perfil dos usuários e que os localiza através de qualquer dispositivo com *Bluetooth* ativo (mais comumente um telefone celular). Para isso, basta o usuário cadastrar o *MAC address* único do celular no aplicativo. Atualmente, a abordagem proposta não contempla a integração com redes sociais, mas é uma boa forma de complementar nosso trabalho, visto que elas detêm algumas informações úteis para personalização.

Em [Blackwell \(2005\)](#), Gerry Blackwell apresenta o Navizon como um sistema de posicionamento *wireless*, que funciona em dispositivos portáteis (e.g. PDA), a partir de *GPS* (quando disponível) e também por triangulação de sinais de pontos de acesso *wireless* e torres de celular GSM. Este sistema oferece aplicações que se baseiam em localização, como encontrar amigos que estejam próximos, disponibilizar pontos de interesse de outros usuários, como restaurantes, teatros entre outros. Todos esses pontos de interesse podem ser marcados com as chamadas *geotags*, que expressam a opinião e avaliação dos usuários.

Em [Kuo et al. \(2009\)](#), é abordado um modelo de recomendação de informações baseado na localização, que se ajusta através do feedback dos usuários que assinam o serviço e também através do teorema de Bayes, para analisar dados históricos de uso.

²<http://www.facebook.com>

Utilizar o feedback como forma de melhorar o perfil é uma boa estratégia, e constitui mais um detalhe que pode ser incorporado à nossa abordagem. Já em [Dao et al. \(2012\)](#), é descrito um modelo de recomendação de publicidade baseada em localização. Este alia filtragem colaborativa, uma técnica frequentemente utilizada para realizar recomendações, em conjunto com algoritmos genéticos, para melhorar a relevância da recomendação. Por fim, em [Bulander et al. \(2005\)](#) é exposto um sistema chamado MoMa (*Mobile Marketing*), que provê publicidade personalizada considerando e respeitando aspectos de privacidade. Nele, os usuários cadastram solicitações de algum serviço e os anunciantes cadastram ofertas. Quando há uma equivalência entre ambos, o usuário é notificado do modo que ele especificou na assinatura do serviço. Outras abordagens não tão relacionadas, mas que buscam a localização como meio de personalização podem ser encontradas em [Rashid et al. \(2008\)](#) e [Aalto et al. \(2004\)](#). A questão de utilizar somente a localização é que fica mais difícil recomendar conteúdo menos intrusivo e mais relevante.

3.2 Discussão

Pelos trabalhos apresentados, podemos perceber que os sistemas evoluíram bastante, passando a não somente utilizar a localização como [Rashid et al. \(2008\)](#), [Blackwell \(2005\)](#) e [Aalto et al. \(2004\)](#), mas também a considerar aspectos de personalização para melhoria do conteúdo provido, seja através do uso de ontologias como em [Liapis et al. \(2008\)](#) e [Panayiotou & Samaras \(2006\)](#) ou até mesmo métodos que auxiliam na tomada de decisões, em [Mazaheri et al. \(2010\)](#) e [Huang & Bian \(2009\)](#).

Em relação à nossa proposta, alguns aspectos podem ser analisados. Diferentemente do que acontece conosco, quase todos os serviços descritos precisam de algum tipo de assinatura ou participação ativa do usuário. Outro aspecto a ser considerado é que podemos utilizar qualquer uma das técnicas de localização mencionadas acima, dado que utilizamos um histórico de localização, em latitude e longitude ou UTM. Esse histórico pode ser obtido, por exemplo, a partir da utilização de um serviço sensível à localização ou de um rastreamento. Com relação à personalização do conteúdo publicitário provido, os dados históricos são cruzados com bases de dados georreferenciadas, permitindo obter algumas informações precisas sobre o perfil do usuário, como escolaridade e renda, atributo este bastante importante ao recomendar anúncios. Devido os dados georreferenciados apresentarem um aspecto probabilístico, utilizamos a mesma estratégia de [Mazaheri et al. \(2010\)](#): redes bayesianas.

Além dos dados georreferenciados cruzados, um conjunto de regras também alimenta a rede bayesiana. Essas regras tratam de um aspecto não identificado em outros trabalhos: a ocupação do usuário. A ocupação é outra variável bastante relevante na identificação do perfil, pois estando de posse dessa informação, é possível refinar alguns atributos. Além disso, temos diferentes particularidades ao tratar com um assalariado ou um estudante, como tipos de roupas usadas, sapatos e valores das ofertas. Por exemplo, dado que o usuário é um estudante, com uma probabilidade de 80%, então podemos restringir o atributo idade a um intervalo mais limitado e recomendar ofertas de cadernos e livros em promoção. Dado todos estes fatores, podemos perceber que nossa proposta traz uma potencial contribuição à área, já que utilizamos certos elementos que até o momento não foram identificados em nenhum trabalho na literatura. Vale destacar que nosso trabalho não foca somente no contexto do transporte público, ele busca enviar anúncios aos usuários de forma geral, seja para aquele que possui carro, que depende de ônibus diariamente ou até mesmo que vai ao trabalho a pé.

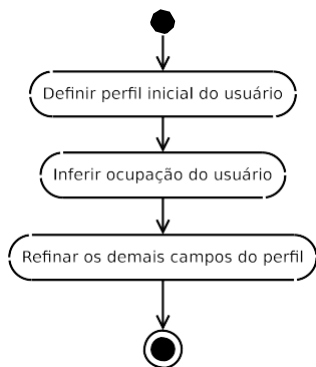
4 Modelo de Criação do Perfil

Neste capítulo será apresentada a proposta do trabalho, descrevendo um serviço do tipo *location-based advertising*, baseado em dados de localização e georreferenciados. Como será visto, nossa maior contribuição é na criação do perfil do usuário, buscando impactar na qualidade da recomendação de ofertas e anúncios ao mesmo. Vale ressaltar que apesar da proposta está condicionada à área de *advertising*, há outros contextos que podem ser pensados, como por exemplo, *e-commerce*. Sites que utilizam mecanismos para fazer recomendação de produtos a usuários poderiam utilizar o perfil criado pela nossa proposta, como uma alternativa, para ajustar os produtos recomendados.

4.1 Workflow da solução

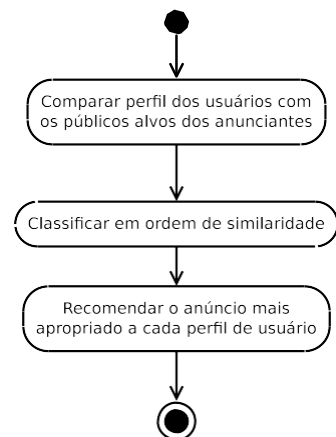
A proposta do nosso trabalho pode ser, fundamentalmente, dividida em duas partes: a criação do perfil probabilístico, com base no histórico de deslocamento de um usuário e a recomendação do anúncio mais adequado a este perfil. Para isto, é necessária uma série de passos, descritos nos diagramas de atividades abaixo (ver Figura 4.1 e Figura 4.2).

Figura 4.1: Diagrama de atividades para criação e refinamento do perfil.



Fonte: Autor, 2012.

Figura 4.2: Diagrama de atividades para realizar a recomendação.



Fonte: Autor, 2012.

As seções a seguir, buscam explicar detalhadamente cada uma das atividades ilustradas acima.

4.2 Definir perfil inicial do usuário

A proposta deste trabalho utiliza como pré-requisito a existência de um **Histórico de Localização** de um dispositivo móvel (usuário) (ilustrado na Figura 4.3), que pode ser gerado por qualquer fonte (e.g. triangulação e/ou *GPS*) e que pode ser obtido por operadoras de celular, sites que possam capturar a localização (Google Maps¹) e também pelo uso de serviços sensíveis à localização (Google Latitude²). Este histórico precisa conter no mínimo as seguintes informações: identificador do dispositivo (número de telefone ou *International Mobile Equipment Identity* (IMEI)) e um conjunto de localizações associadas a informações de data e hora, resultando na tupla [id, localizacao, data/horario], explicada abaixo:

1. **id**: um número de identificação do dispositivo do usuário, geralmente um número de celular;
2. **localizacao**: um ponto de localização espacial, que pode ser formado pelo par ordenado (longitude, latitude) ou em Universal Transverse Mercator (UTM). Esta localização informa exatamente por quais locais o usuário passou;
3. **data/horario**: data e hora em que o usuário estava localizado no local acima.

Estando de posse do histórico, o elemento chamado **Capturador** usa os pontos de localização das tuplas para obter informações georreferenciadas de idade, gênero, escolaridade e renda.

Figura 4.3: Exemplo de um arquivo que mostra o histórico de localização de um usuário durante três dias consecutivos.

```
1 #id_dispositivo, longitude , latitude , dia , mes , ano , hora , minuto
2 dispositivo , -35.7491, -9.6703, 30/01/2012 07:31
3 dispositivo , -35.7059, -9.6475, 30/01/2012 18:00
4 dispositivo , -35.7491, -9.6703, 31/01/2012 07:26
5 dispositivo , -35.7059, -9.6475, 31/01/2012 18:05
6 dispositivo , -35.7491, -9.6703, 01/02/2012 07:45
7 dispositivo , -35.7059, -9.6475, 01/02/2012 18:13
```

Fonte: Autor, 2012.

Mas as informações não são buscadas para qualquer localização, somente para o local aproximado da residência do usuário. Para isto, utiliza-se a primeira localização mais frequente no início da manhã de um dia.

¹<https://maps.google.com/>

²<http://goo.gl/xvKIg>

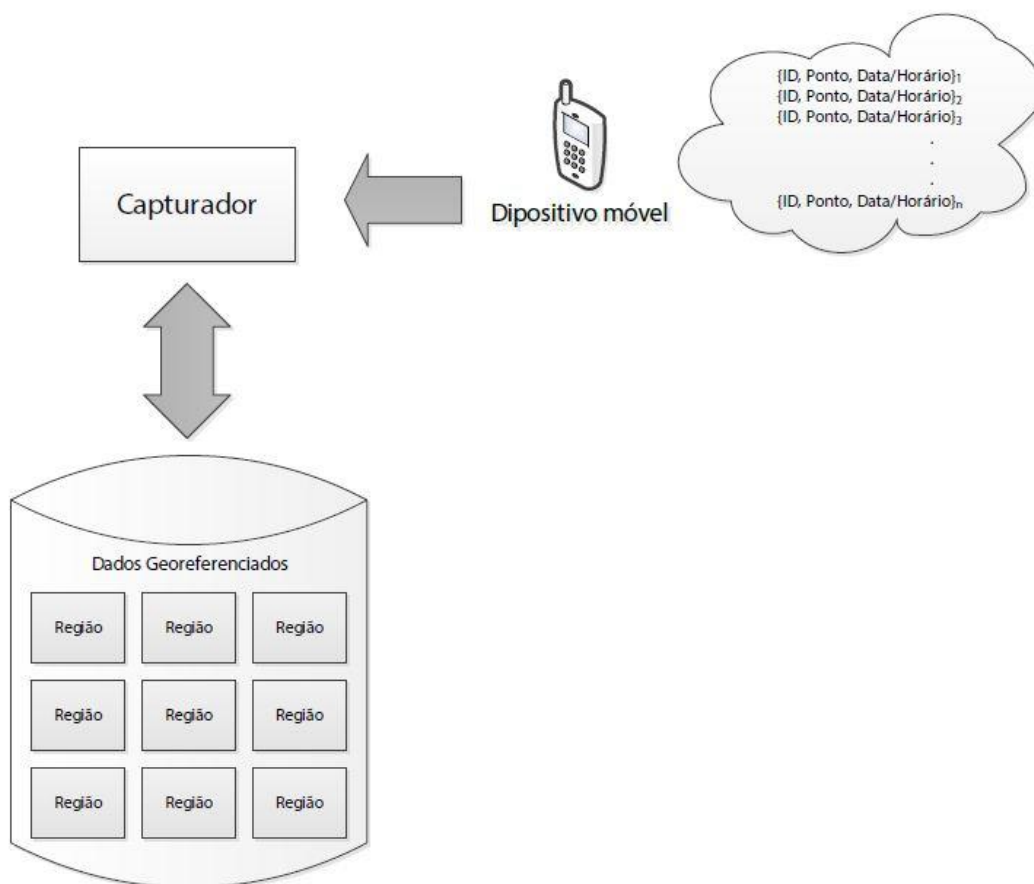
Para ficar mais claro, imaginemos um comportamento geral, para uma pessoa que trabalha e/ou estuda. Normalmente trabalhamos/estudamos durante o dia e à noite, e dormimos durante a madrugada. Em uma situação em que a rotina diária comece cedo, uma pessoa pode acordar às 4h da manhã para ir trabalhar. Neste caso, capturando a localização mais frequente neste horário, podemos inferir onde a mesma mora, e conseqüentemente obtemos as informações associadas àquela localização. Isto foi somente um exemplo, e essa primeira captura pode variar para quem for criar um serviço baseado no trabalho que estamos propondo.

Como mostrada na Figura 4.4, estas informações georreferenciadas são obtidas a partir de bases de dados que descrevem uma população baseado na região. Em alguns casos, um território (e.g. uma cidade, estado ou região) é dividido em partes menores, em que para cada uma são fornecidas um enorme conjunto de informações. Cada uma destas partes pode conter desde uma rua até um ou mais bairros.

Os dados georreferenciados e o histórico são passados a outro elemento chamado **Integrador**, que interage tanto com a **Base de Regras de Classificação** quanto com o **Módulo da Rede Bayesiana**, mostrados na Figura 4.5. Em suma, a base de regras tem por objetivo identificar a ocupação do usuário (ver Seção 4.3). Ela é formada por um conjunto de regras que analisam, através do agrupamento das tuplas por data, horário mais frequente do usuário em determinada região e tempo entre as tuplas do dia, buscando-se inferir se o usuário do dispositivo móvel está mais próximo de um estudante ou de um assalariado, ou até mesmo se ele possui as duas ocupações.

Já o módulo da rede bayesiana destina-se a criar um perfil inicial, com base nas informações georreferenciadas citadas. É importante explicar que este perfil é essencialmente probabilístico, ou seja, ele se baseia em dados de levantamentos feitos para a região. Podemos utilizar um exemplo prático para ilustrar este perfil: dado que um usuário mora no Jacintinho, e que o IBGE, no último censo, registrou que das 2000 pessoas que moram na região, temos 800 homens e 1200 mulheres, podemos inferir que 40% da população é formada por homens enquanto 60% são mulheres. E ainda, se temos a informação que 300 homens possuem o ensino fundamental completo, 400 possuem o ensino médio e 100 possuem o ensino superior, podemos inferir que 50% dos homens possuem o ensino médio completo. Todas estas probabilidades, condicionais ou não, são usadas para preencher a rede bayesiana.

Figura 4.4: Histórico de localização e Base de Dados Georreferenciada.



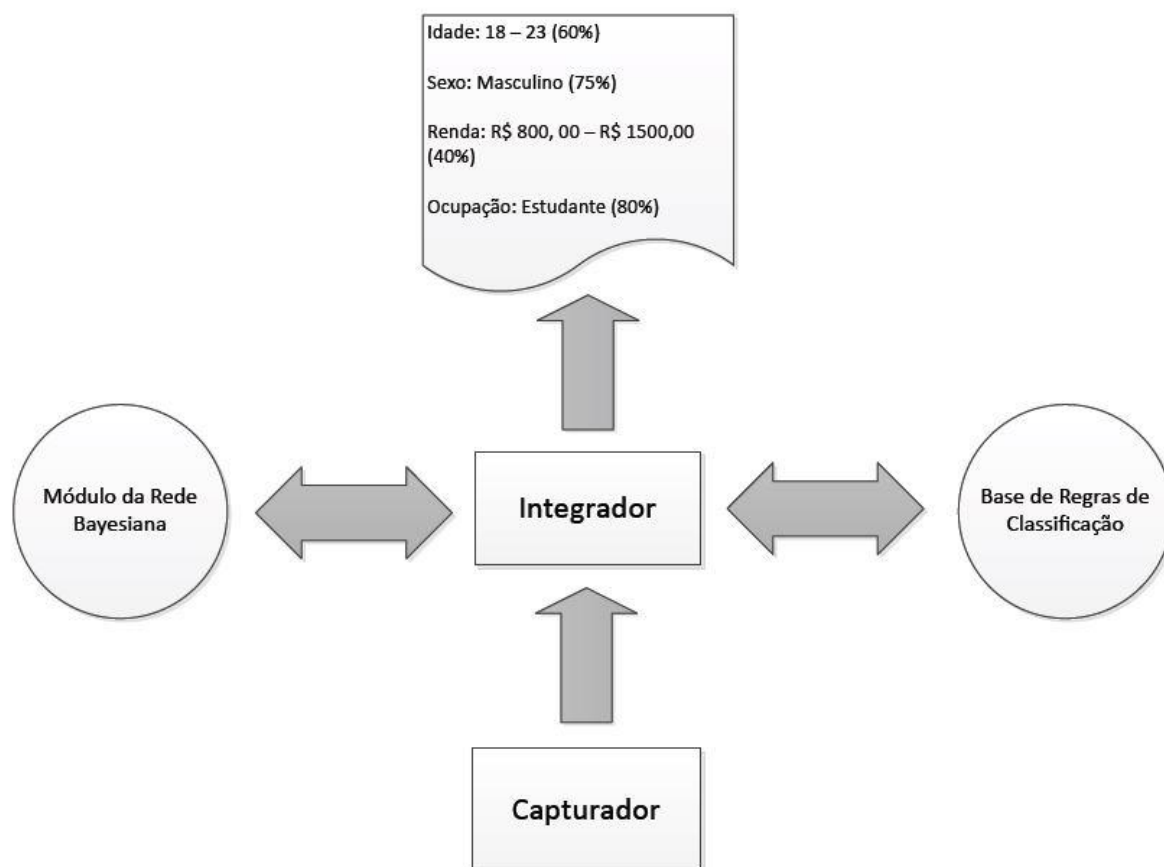
Fonte: Autor, 2012.

O elemento **Módulo da Rede Bayesiana** é composto por uma rede bayesiana, ilustrada na Figura 4.6, que realiza inferências sobre o perfil do usuário, composta pelos nós idade, gênero, escolaridade e renda, além do nó ocupação, sendo este último definido a posteriori pela base de regras. A exemplo da pesquisa desenvolvida por [Yaakub & Napiyah \(2011\)](#) e também com base no modo como é feita a criação de anúncios no Facebook³, idade, gênero, escolaridade e renda são as principais informações buscadas para o alvo de um anúncio. Já no Google Adwords⁴, uma forma de anúncio que vem ganhando espaço, é utilizado somente idade e gênero. Para “alimentar” a rede bayesiana, os dados georreferenciados preenchem as probabilidades a priori de idade, gênero, escolaridade e renda. As probabilidades a priori do nó ocupação são fixas, como descritas na Tabela 4.1. A criação dessa tabela foi baseada em uma pesquisa desenvolvida por [Carvalho & Santos \(2006\)](#), que apresenta um estudo sobre a população e sua mobilidade no ano de 2006.

³<http://www.facebook.com/advertising/>

⁴<https://adwords.google.com/>

Figura 4.5: O Integrador se comunica com a base de regras e o módulo da rede bayesiana, resultando em um perfil probabilístico.



Fonte: Autor, 2012.

Tabela 4.1: Comparação Faixa Etária X Ocupação, para usuários do transporte público coletivo.

Faixas de idade	estudante	assalariado	aposentado
0 - 9	100%	0%	0%
10 - 14	100%	0%	0%
15 - 19	26%	74%	0%
20 - 24	20%	80%	0%
25 - 29	20%	80%	0%
30 - 34	20%	80%	0%
35 - 39	20%	80%	0%
40 - 44	10%	90%	0%
45 - 49	10%	90%	0%
50 - 54	0%	100%	0%
55 - 59	0%	70%	30%
60 - 64	0%	35%	65%
65 - 69	0%	20%	80%
70 - 74	0%	5%	95%
75 - 79	0%	0%	100%
> 80	0%	0%	100%

Fonte: Autor, 2012.

Dado que todas as probabilidades a priori estão preenchidas, a rede é compilada, obtendo uma relação probabilística entre os nós, que será ajustada com a evidência **ocupação**, procedimento que será explicado com mais detalhes na próxima seção. Dessa forma, temos um resultado probabilístico de todos os valores ou faixas de valores assumidos pelos nós, definidos pela base georreferenciada.

Para ilustrar o que foi explicado, considere um usuário que possui um smartphone e que realiza consultas a um sistema que fornece informações sobre os ônibus da cidade onde ele mora. Diariamente, ao sair de casa, ele consulta se o ônibus da linha X está próximo do ponto de ônibus ao qual ele se dirige todas as manhãs. Com a mesma frequência, ao sair do trabalho, ele realiza consultas ao mesmo sistema, com o intuito de voltar para casa. Passado um mês, é obtida uma quantidade de dados suficiente para fazermos as seguintes perguntas: Onde ele mora? Qual o bairro que ele trabalha? Com que frequência o usuário utiliza ônibus? Dado que temos estas informações de consultas ao sistema disponíveis, podemos formar um histórico de localização do usuário. Formado o histórico, recuperamos os dados georreferenciados para o local onde o usuário reside e passamos os mesmos à nossa rede bayesiana, criando um perfil inicial. Ao final do capítulo, será apresentado um exemplo que ajudará a entender todo o processo de inferência do perfil.

4.3 Inferir ocupação do usuário

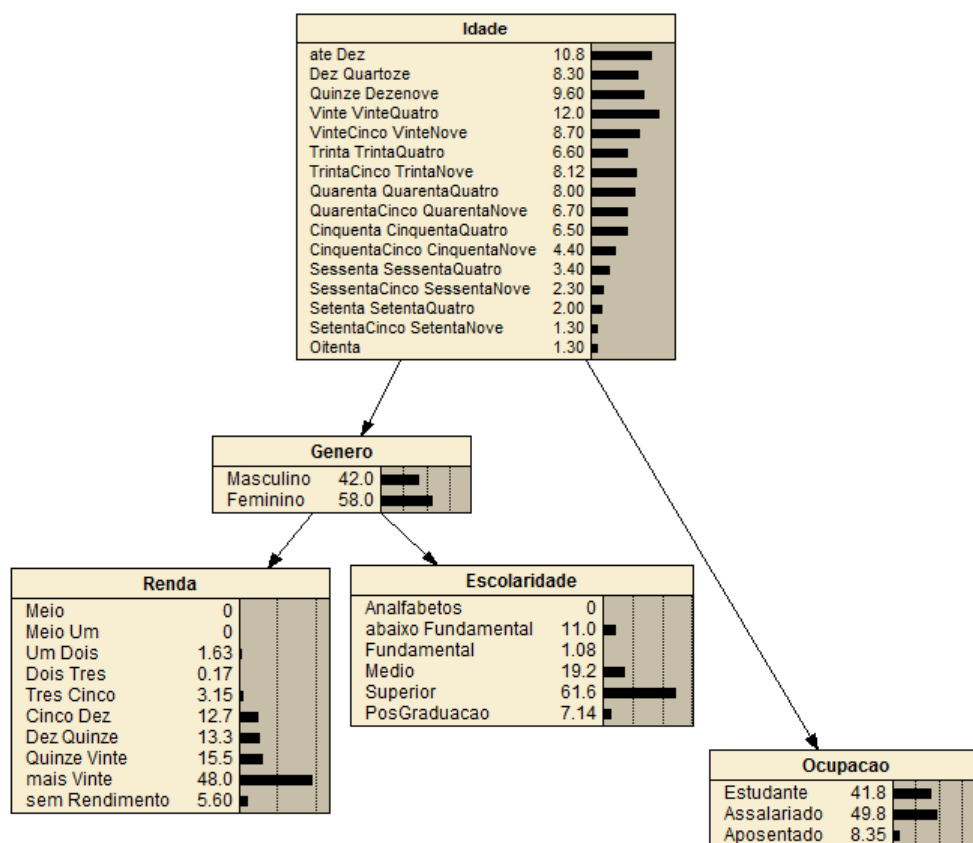
A base de regras, citada na seção anterior, é formada por um conjunto de regras que analisam, através do agrupamento das tuplas por data, horário mais frequente do usuário em determinada região e tempo entre as tuplas do dia, buscando-se inferir se o usuário do dispositivo móvel está mais próximo de um estudante ou de um assalariado, ou até mesmo se ele possui as duas ocupações. Para justificar as ocupações escolhidas, pode-se destacar que em 2006, segundo [Carvalho & Santos \(2006\)](#), estudo e trabalho representavam 65% dos motivos dos deslocamentos nos dias úteis.

Dentre vários métodos que poderiam ser utilizados para “capturar” a ocupação do usuário, decidiu-se escolher uma base de regras devido à simplicidade de implementação e eficiência. Vale mencionar que atualmente o conjunto de regras se limita a definir as ocupações citadas anteriormente, ficando, por hora, em aberto ocupações como autônomo, aposentado entre outras. Devido à quantidade de regras existentes na base, torna-se inviável mostrá-la aqui, podendo ser visto um resumo da mesma na Figura 4.7. É válido destacar que a introdução do nó ocupação no perfil do usuário veio como um diferencial para auxiliar no refinamento do perfil, pois permite o foco em determinadas faixas de idade, permitindo refinar alguns atributos. Quando se trata de anúncios, isto é um aspecto bastante relevante, pois temos diferentes particularidades ao tratar com um assalariado ou um estudante, como o interesse nos tipos de roupas e sapatos e até no poder de compra.

Voltando ao nosso exemplo, vamos supor que temos um usuário que, de segunda a sexta-

feira, realiza consultas à algum sistema que forneça informações sobre os ônibus da cidade. Estas consultas são realizadas pela manhã, aproximadamente às 7h em uma mesma região, e ao final do dia, as consultas acontecem em torno de 18h. Pelas regras apresentadas abaixo (Figura 4.7), vemos que os horários de consulta estão dentro do intervalo de abrangência de *earlyMorning* (pode ser definido com um intervalo de 5h às 8h30) e *earlyNight* (pode ser definido pelo intervalo 18h às 20h). Além disso, o intervalo entre as consultas é maior que 6h, dado o regime de trabalho de um assalariado, nos dias de trabalho (*workday*). Desse modo, as duas regras mostradas vão inferir que o usuário se trata de um assalariado, algo satisfatório e aceitável.

Figura 4.6: Modelo da Rede Bayesiana compilada.



Fonte: Autor, 2012.

Figura 4.7: Exemplo, em pseudocódigo, de algumas regras da Base de Regras.

```

1  regra "earlyMorning"
2      se (hora > 4:55 E hora < 9:01)
3          entao
4              periodo = EARLY_MORNING;
5  fim
6
7  regra "early
Night" 8
9      se (hora > 17:59 E hora < 19:31)
10         entao
11             periodo = EARLY_NIGHT;
12  f
i
m
13
14  regra "earlyMorningWorkDaySaturdayDow
ntown" 15
16     se (periodo == EARLY_MORNING E (dia = WORKDAY || dia = SATURDAY) E intervalo > 6.0)
17         entao
18             ocupacao = HIRED;
19  f
i
m
20
21  regra "earlyNightWorkDayDowntown"
22     se (periodo == EARLY_NIGHT E dia = WORKDAY E intervalo > 6.0)
23         entao
24             ocupacao = HIRED;
25  fim

```

Fonte: Autor, 2012.

4.4 Refinar os demais campos do perfil

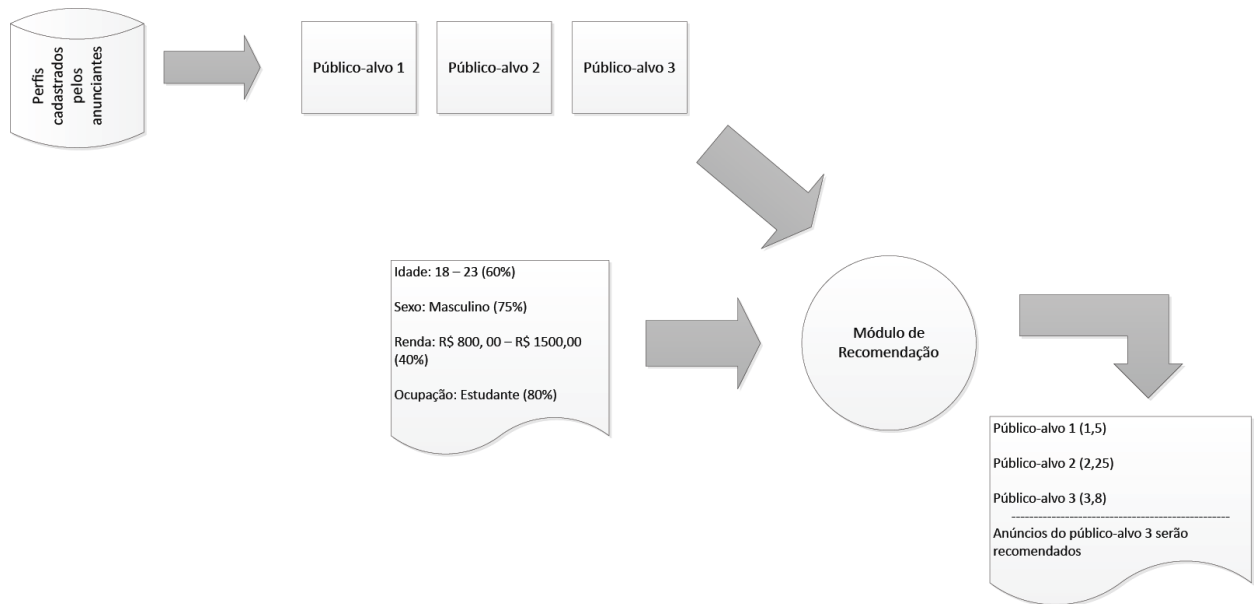
Voltando à Figura 4.5, dado que a base de regras concluiu uma determinada ocupação, a mesma é passada, através do **Integrador**, ao módulo da rede bayesiana, que utiliza a ocupação inferida como uma evidência no nó ocupação. Introduzida esta evidência na rede, ela é recompilada, calculando-se um novo perfil probabilístico do usuário. No caso de um trabalhador assalariado, como é o caso do nosso exemplo, é bem provável que a rede irá alterar a faixa de idade, por exemplo, concentrando as probabilidades nas idades em que a população é economicamente ativa.

4.5 Comparar os perfis dos usuários com os públicos-alvo dos anunciantes

O novo perfil, mencionado na seção anterior e ilustrado na Figura 4.8, será comparado através de métricas de similaridade, no **Módulo de Recomendação**, com os perfis dos públicos-alvo, cadastrados pelos anunciantes. Estes perfis são cadastrados através de uma página Web, na qual o anunciante além de informar o perfil de interesse do público-alvo também cadastra sua empresa, informando nome, contato, endereço entre outras coisas.

É importante mencionar que a similaridade entre anunciante e usuário é resultante da soma das similaridades de cada um dos cinco atributos, valor que varia de 0,0 a 5,0. Para isto, cada atributo dos perfis é representado no formato de um vetor. Em seguida, é calculada e armazenada a similaridade entre os dois vetores daquele atributo, pois futuramente este valor será somado aos demais. Inclusive é possível ponderar a similaridade, sendo estabelecida de acordo com a importância dos atributos a ser julgada pelos anunciantes.

Figura 4.8: O perfil probabilístico é comparado aos perfis dos públicos-alvo cadastrados pelos anunciantes.



Fonte: Autor, 2012.

4.6 Classificar em ordem de similaridade

Ainda no elemento **Módulo de Recomendação**, temos a criação de uma lista, em ordem decrescente, da comparação do perfil probabilístico do usuário com todos os perfis cadastrados na base de dados pelos anunciantes. Ao final, é eleito o perfil mais similar a ser recomendado à aquele usuário.

Retomando nosso exemplo do usuário assalariado, e supondo que temos em nossa base de dados o cadastro de uma loja de tênis, um supermercado, um salão de beleza e um centro de estética, o perfil mais adequado ao perfil gerado será escolhido. Neste caso, possivelmente teremos ou um supermercado ou uma loja de tênis.

4.7 Recomendar o anúncio mais apropriado a cada perfil de usuário

Como saída de todo esse conjunto de atividades, temos qual oferta será enviada ao usuário, ilustrado na Figura 4.8. Atualmente, o modelo implementado recomenda somente o anúncio/oferta daquele anunciante que mais se aproxima ao perfil do cliente, priorizando a qualidade máxima da recomendação. E com uma simples mudança, poderíamos facilmente entregar ao cliente os dez anúncios que foram eleitos os mais adequados. Um outro direcionamento, que inverte a lógica explicada anteriormente e que traz resultados relevantes é: dado um conjunto de clientes, quais são os **N** mais adequados a determinado perfil cadastrado pelo anunciante?

4.8 Exemplo passo-a-passo

Para deixar mais claro tudo o que foi explicado anteriormente sobre a proposta, vamos utilizar um exemplo bem simples, no qual explicaremos o passo-a-passo, desde a obtenção do histórico até o envio da oferta. Para enviarmos uma oferta a um determinado usuário, primeiramente necessitamos de um histórico de localização do mesmo. Para o nosso exemplo, vamos supor que uma pessoa utilizou um sistema de previsão de chegada de ônibus na cidade em que mora, durante uma semana. O sistema monitorou o uso do usuário, que o utilizava todos os dias, ao sair de casa em direção ao ponto e também ao sair do trabalho. Assim, foi obtido o seguinte histórico:

A partir do histórico mostrado na Figura 4.9, a primeira coisa que precisamos inferir é a região onde o usuário reside. Para esta situação, vamos considerar aquela região mais frequente na qual o usuário faz o primeiro uso do sistema. As informações de latitude e longitude nos indicam que a região de residência fica nas proximidades do bairro da Mangabeiras. Com isto, vamos utilizar a base de dados georreferenciada para obter as informações de idade, gênero, renda e escolaridade. O tipo de informação fornecida pela base de dados é da seguinte forma

Figura 4.9: Histórico de localização obtido pelo uso do sistema de previsão de chegada.

	#id	longitude	latitude	dia	mes	ano	hora	minuto
1	914	-35.7068	-9.6428	02/04/2012			06:59	
3	914	-35.7193	-9.6719	02/04/2012			18:24	
4	914	-35.7070	-9.6423	03/04/2012			07:04	
5	914	-35.7199	-9.6717	03/04/2012			18:46	
6	914	-35.7063	-9.6427	04/04/2012			07:17	
7	914	-35.7202	-9.6713	04/04/2012			18:41	
8	914	-35.7069	-9.6424	05/04/2012			07:06	
9	914	-35.7197	-9.6715	05/04/2012			18:45	
10	914	-35.7060	-9.6426	06/04/2012			07:13	
11	914	-35.7205	-9.6717	06/04/2012			18:22	

Fonte: Autor, 2012.

(para o atributo idade):

Pessoas com até 10 anos de idade: 50

Pessoas de 10 a 14 anos de idade: 35

Pessoas de 15 a 19 anos de idade: 41

Pessoas de 20 a 24 anos de idade: 73

Pessoas de 25 a 29 anos de idade: 29

Pessoas de 30 a 34 anos de idade: 21 e assim por diante...

Todas estas informações serão usadas para preencher as tabelas de probabilidade da rede bayesiana, como mostrado na Figura 4.10. Quando as tabelas de todos os atributos estiverem preenchidos, a rede é compilada, resultando no que é mostrado na Figura 4.11. Neste momento, precisamos inferir a ocupação do usuário para ajustar a rede. Fornecendo como entrada para a base de regras o histórico de localização, dentre as várias regras que formam a base, aquelas mostradas na Figura 4.12 serão executadas, o que permitirá inferir que o usuário se trata de um assalariado com 100% de probabilidade.

Tendo esta nova informação, ela é inserida na rede no formato de uma nova evidência, obtendo a rede mostrada na Figura 4.13. Uma vez obtido este perfil probabilístico do usuário, já com a evidência da ocupação, a próxima etapa é criar um vetor para cada atributo, que represente o perfil:

Idade < 0 ; 0 ; 0,0482 ; 0,0963 ... >

Gênero < 0,416 ; 0,584 >

Escolaridade < 0 ; 0,11 ; 0,018 ; 0,193 >

Renda < 0 ; 0 ; 0,0163 ; 0,0017 ... >

Ocupação < 0 ; 1 ; 0 >

Figura 4.10: Exemplo de uma tabela de probabilidade condicional para o atributo gênero, dado a idade.

Idade	Masculino	Feminino
ate Dez	50	50
Dez Quatoze	50.5	49.5
Quinze Dezenove	41	59
Vinte VinteQuatro	37	63
VinteCinco VinteNove	41	59
Trinta TrintaQuatro	29	71
TrintaCinco TrintaNove	36	64
Quarenta QuarentaQuatro	42	58
QuarentaCinco QuarentaNove	45	55
Cinquenta CinquentaQuatro	49	51
CinquentaCinco CinquentaNove	52	48
Sessenta SessentaQuatro	51	49
SessentaCinco SessentaNove	30	70
Setenta SetentaQuatro	33	67
SetentaCinco SetentaNove	28	72
Oitenta	29	71

Fonte: Autor, 2012.

Esta é a representação do usuário. Para cada vetor, cada um de seus componentes pode variar de 0,0 a 1,0 (probabilidade de cada valor).

Tendo de antemão o público-alvo dos anunciantes (ver Tabela 4.2), criamos, de modo análogo, os vetores que representem cada um. Neste caso, todo valor que fizer parte do público-alvo é preenchido com 1,0 e para os outros é 0,0, ou seja no caso do Anunciante 1, como o interesse é em ambos os gêneros, seu vetor gênero fica $\langle 1,0 ; 1,0 \rangle$. Já para o Anunciante 2, temos o vetor $\langle 1,0 ; 0,0 \rangle$. De modo análogo, os demais vetores são construídos. No caso de idade e renda, que são formados por intervalos, o valor é calculado proporcionalmente.

Tabela 4.2: Tabela de públicos-alvo dos anunciantes

Anunciante	Idade	Gênero	Renda	Escolaridade	Ocupação
Anunciante 1	18 - 57	Ambos	1.000 - 5.000	Superior	Estudante
Anunciante 2	05 - 75	Masculino	4.000 - 100.000	Todos	Ambos
Anunciante 3	20 - 40	Feminino	2.000 - 10.000	Todos	Assalariado

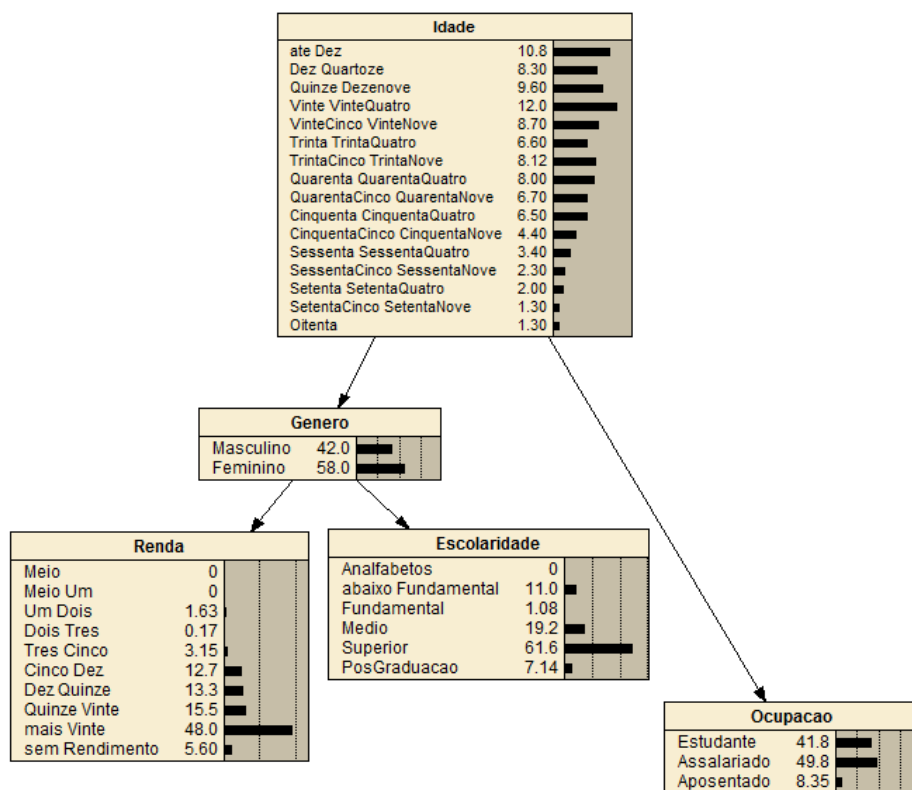
Fonte: Autor, 2012.

Tendo os vetores para cada público-alvo, basta agora calcularmos a similaridade com o perfil gerado para o usuário. A similaridade é resultado da soma das similaridades de cada vetor, resultando em 5,0 no máximo (cinco atributos, que resultam em cinco vetores). Para dar

mais detalhes desta etapa, vamos calcular a similaridade entre os vetores gênero e ocupação do usuário e do Anunciante 1 (consultar a Seção 2.4 para eventuais dúvidas):

Vetor gênero do usuário (v1): < 0,416 ; 0,584 >

Figura 4.11: Exemplo de uma rede bayesiana compilada após todas as tabelas terem sido preenchidas.



Fonte: Autor, 2012.

Figura 4.12: Regras que são usadas para inferir a ocupação para o usuário do nosso

```

1  regra "earlyMorning"
2    se (hora > 4:55 E hora < 9:01)
3      entao
4        periodo = EARLY_MORNING;
5  fim
6
7  regra "earlyNight" 8
9    se (hora > 17:59 E hora < 19:31)
10   entao
11     periodo = EARLY_NIGHT;
12  fim
13
14  regra "earlyMorningWorkDaySaturdayDowntown" 15
16    se (periodo == EARLY_MORNING E (dia = WORKDAY || dia = SATURDAY) E intervalo > 6.0)
17      entao
18        ocupacao = HIRED;
19  fim
20
21  regra "earlyNightWorkDayDowntown"
22    se (periodo == EARLY_NIGHT E dia = WORKDAY E intervalo > 6.0)
23      entao
24        ocupacao = HIRED;
25  fim
    
```

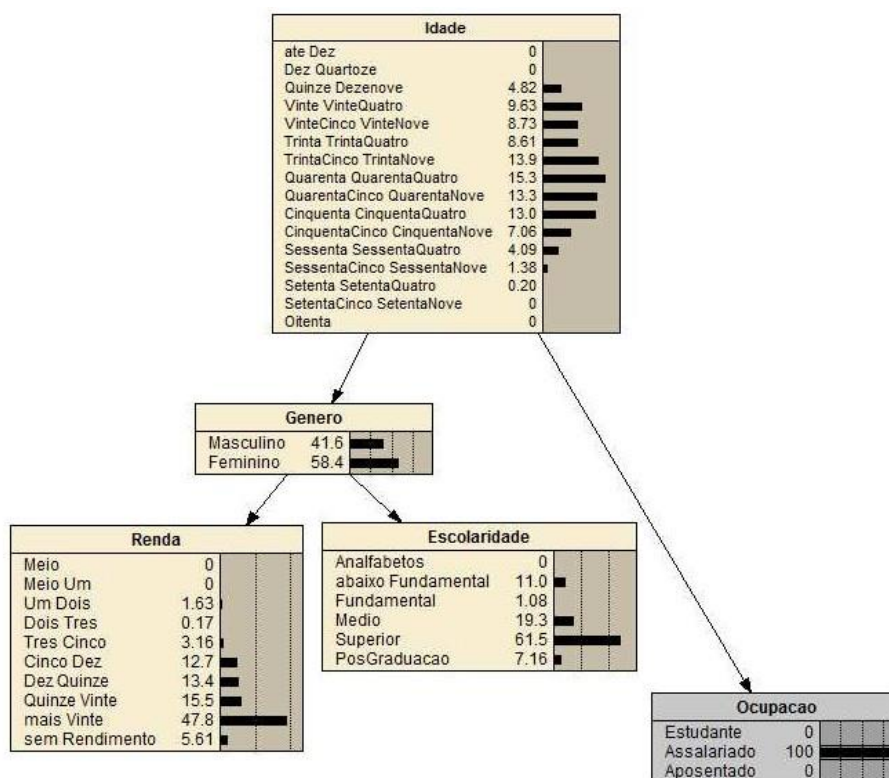
Fonte: Autor, 2012.

Vetor gênero do Anunciante 1 (v2): $\langle 1,0 ; 1,0 \rangle$

Produto interno: $0,416 * 1 + 0,584 * 1 = 1,0$

Norma de v1: $0,416 * 0,416 + 0,584 * 0,584 = 0,341$

Figura 4.13: Rede bayesiana com a evidência de ocupação inserida.



Fonte: Autor, 2012.

Norma de v2: $1,0 * 1,0 + 1,0 * 1,0 = 2,0$

Similaridade: $1,0 / (\sqrt{0,341} * \sqrt{2,0}) = 0,826$

Vetor ocupação do usuário (v3): $\langle 0,0 ; 1,0 ; 0,0 \rangle$

Vetor ocupação do Anunciante 1 (v4): $\langle 1,0 ; 0,0 ; 0,0 \rangle$

Produto interno: $0,0 * 1,0 + 1,0 * 0,0 + 0,0 * 0,0 = 0,0$

Norma de v3: $0,0 * 0,0 + 1,0 * 1,0 + 0,0 * 0,0 = 1,0$

Norma de v4: $1,0 * 1,0 + 0,0 * 0,0 + 0,0 * 0,0 = 1,0$

Similaridade: $0,0 / (\sqrt{1,0} * \sqrt{1,0}) = 0,0$

Da mesma forma, são calculadas as similaridades para os demais vetores, e ao fim, somamos tudo. Fazendo o restante dos cálculos, vamos obter algo próximo das similaridades abaixo:

Anunciante 1: 2,9

Anunciante 2: 4,2

Anunciante 3: 3,4

Ordenando de maneira decrescente, percebemos que o público-alvo do Anunciante 2 é o que mais se aproxima do usuário em questão. Conseqüentemente, anúncios do Anunciante 2 serão enviados ao usuário.

4.9 Limitações

Pode-se perceber na descrição da proposta e no exemplo dado na seção anterior, que nosso trabalho possui algumas limitações. Primeiramente, um pré-requisito é a existência de um histórico ou conjunto de históricos de localização para dar início a todo o processo de inferência do perfil, que deve ser obtido através de técnicas externas à proposta. Outro ponto é a existência de uma base de dados georreferenciada íntegra, consistente e confiável para buscar as informações da população de uma região de interesse.

Outra limitação é que nossa base de regras se limita a inferir se um usuário se trata de um estudante, assalariado ou ambos, deixando em aberto às demais ocupações que uma pessoa pode assumir. Isto ocorre em virtude da complexidade que está relacionada a criar um padrão de deslocamento para um aposentado, um autônomo ou até mesmo um desempregado. Por fim, para enviar o(s) anúncio(s) ao usuário, precisamos de serviços de terceiros que possuam uma infraestrutura para prover esta capacidade.

4.10 Implementação da proposta

Partindo da ideia conceitual do trabalho, só foi possível obter e avaliar os resultados devido à implementação realizada. A seguir serão mostrados alguns trechos de código das principais partes do trabalho. A Figura 4.14 mostra uma parte da classe responsável por instanciar e inserir os dados na rede bayesiana. Para isto, foi usado a *Application programming interface* (API) do UnBBayes⁵.

⁵<http://unbbayes.sourceforge.net/>

Figura 4.14: Classe que instancia e alimenta a rede bayesiana.

```

1 public class BayesNetworkIBGE extends BayesNetwork{
2
3     public BayesNetworkIBGE(String name, String path) throws IOException{
4         super(name, path);5
5     }
6
7     @Override
8     public List<ProbabilisticNode> updateTable(Map<String, Float> age, float [][] gender,
9         float [][] income, float [][] education) throws IOException{
10        List<ProbabilisticNode> probabilisticNodeList = new ArrayList<ProbabilisticNode>();
11
12        ProbabilisticNode ageNode = (ProbabilisticNode) network.getNodeAt(0);
13        ProbabilisticNode genderNode = (ProbabilisticNode) network.getNodeAt(1);
14        ProbabilisticNode incomeNode = (ProbabilisticNode) network.getNodeAt(2);
15        ProbabilisticNode educationNode = (ProbabilisticNode) network.getNodeAt(3);
16        ProbabilisticNode occupationNode = (ProbabilisticNode) network.getNodeAt(4);
17
18        PotentialTable ageTable = (PotentialTable) ageNode.getProbabilityFunction();
19        PotentialTable genderTable = (PotentialTable) genderNode.getProbabilityFunction();
20        PotentialTable incomeTable = (PotentialTable) incomeNode.getProbabilityFunction();
21        PotentialTable educationTable = (PotentialTable) educationNode.getProbabilityFunction();
22        PotentialTable occupationTable = (PotentialTable) occupationNode.getProbabilityFunction();
23
24        for(int i = 0; i <= DefaultGisDatabase.SIZE_MAP.get("age") - 1; i++)
25            ageTable.setValue(i, age.get(ageNode.getStateAt(i)));
26        probabilisticNodeList.add(ageNode);
27
28        for(int i = 0; i <= DefaultGisDatabase.SIZE_MAP.get("age") - 1; i++)
29            for(int j = 0; j <= DefaultGisDatabase.SIZE_MAP.get("gender") - 1; j++)
30                genderTable.setValue(i*2+j, gender[i][j]);
31        probabilisticNodeList.add(genderNode);
32
33        for(int i = 0; i <= DefaultGisDatabase.SIZE_MAP.get("gender") - 1; i++)
34            for(int j = 0; j <= DefaultGisDatabase.SIZE_MAP.get("income") - 1; j++)
35                incomeTable.setValue(i*10+j, income[i][j]);
36        probabilisticNodeList.add(incomeNode);
37
38        for(int i = 0; i <= DefaultGisDatabase.SIZE_MAP.get("gender") - 1; i++)
39            for(int j = 0; j <= DefaultGisDatabase.SIZE_MAP.get("education") - 1; j++)
40                educationTable.setValue(i*6+j, education[i][j]);
41        probabilisticNodeList.add(educationNode);
42
43        for(int i = 0; i <= DefaultGisDatabase.SIZE_MAP.get("age") - 1; i++)
44            for(int j = 0; j <= DefaultGisDatabase.SIZE_MAP.get("occupation") - 1; j++)
45                occupationTable.setValue(i*3+j, OCCUPATION_APRIORI[i*3+j]);
46        probabilisticNodeList.add(occupationNode);
47
48        BaseIO io = new NetIO();
49        try {
50            network.compile();
51        } catch (Exception ex) {
52            Logger.getLogger(BayesNetworkIBGE.class.getName()).log(Level.SEVERE, null, ex);53
53        }
54        io.save(new File(path), network);
55
56        return Collections.unmodifiableList(probabilisticNodeList);
57    }
58
59    ...
60
61 }

```

Fonte: Autor, 2012.

Já a Figura 4.15 mostra, parcialmente, a classe responsável por gerenciar e executar as regras de inferência de ocupação exibidas na Figura 4.16. Para isso, utilizou-se a API do Drools⁶, uma sistema de gerenciamento de regras de negócio.

⁶<http://www.jboss.org/drools/>

Figura 4.15: Classe que gerencia e executa a base de regras responsáveis pela inferência da ocupação.

```
1 public class Drools {
2
3     public void execute(DataSet dt) {
4         List<LocalDate> ld = dt.listOfDates();
5         LocalDate localDate;
6
7         try {
8             // load up the knowledge base
9             KnowledgeBase kbase = readKnowledgeBase();
10            StatefulKnowledgeSession ksession = kbase.newStatefulKnowledgeSession();
11            KnowledgeRuntimeLogger logger = KnowledgeRuntimeLoggerFactory.newFileLogger(ksession, "InferenceRules");
12            // go !
13            List<MobileData> treatedDataset = dt.treatDatasetToDecisionTree();
14
15            ksession.insert(dt);
16
17            for(int i = 0; i <= ld.size() - 1; i++){
18                localDate = ld.get(i);
19
20                for (MobileData md : treatedDataset) {
21
22                    if(md.getDate().toLocalDate().equals(localDate)){
23                        ksession.insert(md);
24                    }
25                }
26
27                ksession.fireAllRules();
28
29                dt.cleanAnalysedPeriod();
30            }
31
32            logger.close();
33
34        } catch (Throwable t) {
35            t.printStackTrace();
36        }
37
38    }
39
40    ...
41
42 }
```

Fonte: Autor, 2012.

Figura 4.16: Exemplo de algumas regras da Base de Regras.

```

1  rule "daybreak"
2      when
3          c : MobileData( eval(c.isAfter(23, 59) && c.isBefore(4, 40)), period == null );
4      then
5          c.setPeriod(Period.DAYBREAK);
6          System.out.println(c.getPeriod());
7          update (c);
8  end
9
10 rule "earlyMorning"
11     when
12         c : MobileData( eval(c.isAfter(4, 55) && c.isBefore(9, 1)), period == null );
13     then
14         c.setPeriod(Period.EARLY_MORNING);
15         System.out.println(c.getPeriod());
16         update (c);
17 end
18
19 rule "middleAfternoon"
20     when
21         c : MobileData( eval(c.isAfter(14, 10) && c.isBefore(16, 55)), period == null );
22     then
23         c.setPeriod(Period.MIDDLE_AFTERNOON);
24         System.out.println(c.getPeriod());
25         update (c);
26 end
27
28 rule "earlyNight"
29     when
30         c : MobileData( eval(c.isAfter(17, 59) && c.isBefore(19, 31)), period == null );
31     then
32         c.setPeriod(Period.EARLY_NIGHT);
33         System.out.println(c.getPeriod());
34         update (c);
35 end
36
37 rule "earlyMorningWorkDaySaturdayDowntown"
38     when
39         c : MobileData(period == Period.EARLY_MORNING, eval(c.isWorkDay() || c.isSaturday()));
40         dt : DataSet( eval(dt.intervalsTimeByDate(c.getDate()) > 6.0) )
41     then
42         dt.addOccupation(Occupation.HIRED);
43         dt.addAnalysedPeriod(Period.EARLY_MORNING);
44         dt.addAnalysedDay(c.getDate());
45         System.out.println(Occupation.HIRED);
46     end
47
48 rule "earlyNightWorkDayDowntown"
49     when
50         c : MobileData(period == Period.EARLY_NIGHT, eval(c.isWorkDay()));
51         dt : DataSet( eval(dt.intervalsTimeByDate(c.getDate()) > 6.0) )
52     then
53         dt.addOccupation(Occupation.HIRED);
54         dt.addAnalysedPeriod(Period.EARLY_NIGHT);
55         dt.addAnalysedDay(c.getDate());
56         System.out.println(Occupation.HIRED);
57     end
58
59 rule "earlyAfternoonWorkDaySchool"
60     when
61         c : MobileData(period == Period.EARLY_AFTERNOON, eval(c.isWorkDay()));
62         dt : DataSet( eval(dt.intervalsTimeByDate(c.getDate()) >= 4.0 && dt.intervalsTimeByDate(c.getDate()) <= 6.8) )
63         eval(!dt.getAnalysedPeriods().contains(Period.EARLY_MORNING))
64     then
65         dt.addOccupation(Occupation.STUDENT);
66         dt.addAnalysedPeriod(Period.EARLY_AFTERNOON);
67         dt.addAnalysedDay(c.getDate());
68         System.out.println(Occupation.STUDENT);
69     end
70
71 rule "endMorningEarlyMorningQuery"
72     when
73         c : MobileData(period == Period.END_MORNING || period == Period.EARLY_AFTERNOON,
74             eval(c.isWorkDay()));
75         dt : DataSet( eval(dt.intervalsTimeByDate(c.getDate()) >= 4.0 && dt.intervalsTimeByDate(c.getDate()) <= 6.8) )
76         eval(dt.getAnalysedPeriods().contains(Period.EARLY_MORNING))
77     then
78         dt.addOccupation(Occupation.STUDENT);
79         dt.addAnalysedPeriod(Period.END_MORNING);
80         dt.addAnalysedDay(c.getDate());
81         System.out.println(Occupation.STUDENT);
82     end
83
84 end

```

Por fim, temos as Figuras 4.17 e 4.18, que detalham o método-chave para o cálculo da similaridade para a abordagem proposta pelo trabalho. Para isso, recupera-se da base de dados o público-alvo de um anunciante, cria-se o(s) vetor(es) do(s) perfil(is) dos usuários utilizando a abordagem proposta, invoca-se o método responsável pelo cálculo da similaridade e, ao final, retorna-se todas as similaridades entre cada usuário e o anunciante solicitado.

Figura 4.17: Método que compara o público-alvo de cada anunciante com um perfil de usuário gerado pela proposta do trabalho (Parte 1).

```

1  public static Set<CosineMap> doInferenceByAdvertiserFromIBGE (String brand) throws IOException, Exception {
2      BayesNetworkIBGE bn = loadBayesianNetwork ("BayesNetwork", NETWORK_LOCATION);
3      List<ProbabilisticNode> probabilisticNodeList;
4      String listCS;
5
6      SortedSet<CosineMap> listOfCustomers = new TreeSet<CosineMap>();
7
8      DataSet dataset;
9
10     Configuration cfg = new Configuration().configure("hibernate-lba.cfg.xml");
11     HibernateUtil.getSessionFactory(cfg);
12     HibernateUtil.getCurrentSession().beginTransaction();
13
14     AdvertiserHibernateDAO advertisers = new AdvertiserHibernateDAO();
15     List<Advertiser> advertiserList = advertisers.findAll();
16
17     Map<String, ProfileVector> advertisersMap =
18         generateAdvertisersProfileVectors(advertiserList);
19
20     if(!advertisersMap.keySet().contains(brand)){
21         System.out.println("There isn't the brand " + brand);
22         return null;
23     }
24
25     File[] files = new File(FINAL_DATASET_LOCATION).listFiles();
26     ProfileVector customerVector;
27
28     for(int j = 1; j < files.length; j++) {
29         dataset = Utils.loadDataset(files[j].getPath());
30         listCS = dataset.getHomeSectors();
31
32         if(listCS == "") {
33             listOfCustomers.add(new CosineMap(files[j].getPath(), 0));
34             continue;
35         }
36         if(listCS == null){
37             System.out.println("Setor censitário não encontrado!");
38             System.out.println("usuário " + files[j].getPath());
39             listOfCustomers.add(new CosineMap(files[j].getPath(), 0));
40             continue;
41         }
42
43         customerVector = new ProfileVector();
44
45         probabilisticNodeList = bn.updateTable(gdb.inferAge(listCS),
46             gdb.inferGender(listCS),
47             gdb.inferIncome(listCS),
48             gdb.inferEducation(listCS));
49
50         ProbabilisticNode ageNode = probabilisticNodeList.get(0);
51         ProbabilisticNode genderNode = probabilisticNodeList.get(1);
52         ProbabilisticNode incomeNode = probabilisticNodeList.get(2);
53         ProbabilisticNode educationNode = probabilisticNodeList.get(3);
54         ProbabilisticNode occupationNode = probabilisticNodeList.get(4);
55
56         Map<Occupation, Float> occupation = DecisionTree.inferOccupation(dataset);
57
58         float[] likelihoodOccupation = new float[DefaultGisDatabase.SIZE_MAP.get("occupation")];
59         if(occupation.isEmpty() || occupation.containsKey(Occupation.UNDEFINED)) {
60             System.out.println("Não é possível inferir a ocupação para " + files[j].getPath());
61             customerVector = doInferenceByAdvertiserFromIBGENoOccupation(listCS);
62         }
63     }
64     ...

```

Fonte: Autor, 2012.

Figura 4.18: Método que compara o público-alvo de cada anunciante com um perfil de usuário gerado pela proposta do trabalho (Parte 2).

```

98     ...
99     else {
100         Map<String, Float> clone = bn.translateToBayesNet(occupation);
101
102         for (int i = 0; i < DefaultGisDatabase.SIZE_MAP.get("occupation"); i++) {
103             if (clone.containsKey(occupationNode.getStateAt(i))) {
104                 likelihoodOccupation[i] = clone.get(occupationNode.getStateAt(i));
105             } else {
106                 likelihoodOccupation[i] = 0.00f;
107             }
108         }
109         System.err.println("\n_ocupacao_");
110         System.err.println(Arrays.toString(likelihoodOccupation));
111         occupationNode.addLikelihood(likelihoodOccupation);
112         bn.getNetwork().updateEvidences();
113         customerVector.setOccupation(likelihoodOccupation);
114
115         double[] aux;
116         aux = new double[DefaultGisDatabase.SIZE_MAP.get("age")];
117         for (int i = 0; i < aux.length; i++) {
118             aux[i] = ageNode.getMarginalAt(i);
119         }
120         customerVector.setAge(aux);
121         System.err.println("\n_idade_");
122         System.err.println(Arrays.toString(aux));
123
124         aux = new double[DefaultGisDatabase.SIZE_MAP.get("gender")];
125         for (int i = 0; i < aux.length; i++) {
126             aux[i] = genderNode.getMarginalAt(i);
127         }
128         customerVector.setGender(aux);
129         System.err.println("\n_genero_");
130         System.err.println(Arrays.toString(aux));
131         aux = new double[DefaultGisDatabase.SIZE_MAP.get("income")];
132         for (int i = 0; i < aux.length; i++) {
133             aux[i] = incomeNode.getMarginalAt(i);
134         }
135         customerVector.setIncome(aux);
136         System.err.println("\n_renda_");
137         System.err.println(Arrays.toString(aux));
138
139         aux = new double[DefaultGisDatabase.SIZE_MAP.get("education")];
140         for (int i = 0; i < aux.length; i++) {
141             aux[i] = educationNode.getMarginalAt(i);
142         }
143         customerVector.setEducation(aux);
144         System.err.println("\n_escolaridade_");
145         System.err.println(Arrays.toString(aux));
146     }
147 }
148
149 listOfCustomers.add(new CosineMap(files[j].getPath(),
150     Utils.cosine(customerVector.getAge(), advertisersMap.get(brand).getAge())
151     + Utils.cosine(customerVector.getGender(), advertisersMap.get(brand).getGender())
152     + Utils.cosine(customerVector.getIncome(), advertisersMap.get(brand).getIncome())
153     + Utils.cosine(customerVector.getEducation(), advertisersMap.get(brand).getEducation())
154     + Utils.cosine(customerVector.getOccupation(), advertisersMap.get(brand).getOccupation())));
155
156
157 }
158
159 return listOfCustomers;
160 }

```

Fonte: Autor, 2012.

5 Descrição do estudo de caso

Avaliar novas abordagens e técnicas é uma das atividades fundamentais em qualquer área, principalmente na pesquisa acadêmica. O desenvolvimento de estudos de caso leva a este caminho da experimentação. Assim, neste capítulo daremos uma introdução ao estudo de caso que foi realizado em ambiente real, abordado no Capítulo 6, e contextualizaremos um método chamado Goal/Question/Metric (GQM), escolhido para estruturá-lo.

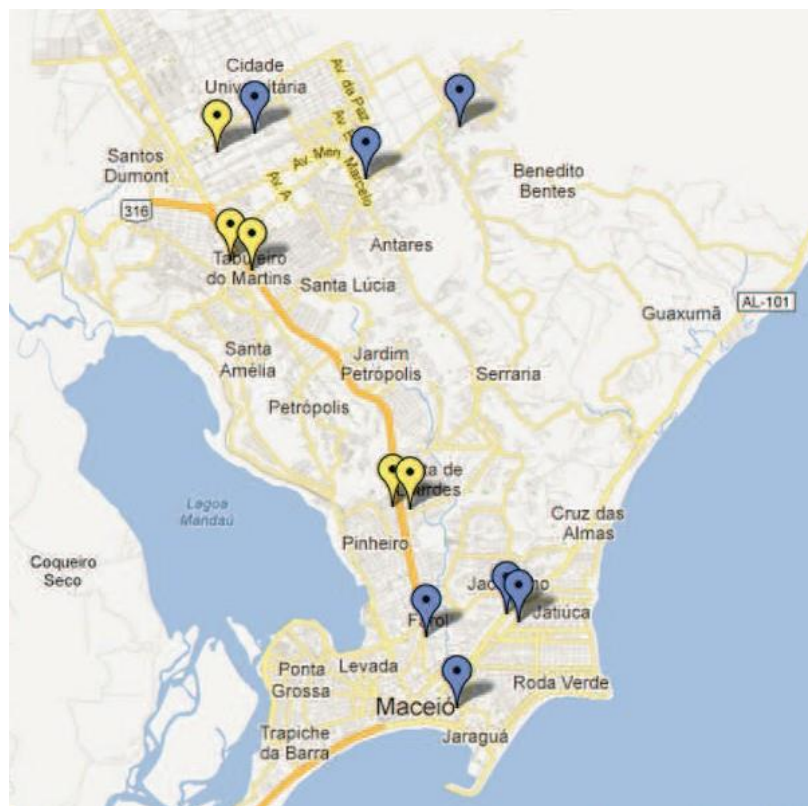
5.1 Descrição geral da metodologia utilizada

Para realizar nosso estudo de caso, inicialmente 41 pessoas, que dependem do transporte público coletivo, foram entrevistadas em alguns pontos de ônibus da cidade de Maceió, nos horários de maior movimento. Estudantes, assalariados, idosos, autônomos e desempregados foram submetidos a um questionário que contém uma série de campos necessários para criar além do perfil pessoal, o perfil do mesmo como usuário de transporte coletivo, composto, por exemplo, pela frequência que ele utiliza ônibus, os bairros e horários onde o usuário costuma pegar os ônibus e as linhas que ele costuma utilizar (ver Apêndice A). Após a aplicação do questionário, os resultados foram organizados em uma planilha, e serviram como base para a criação de 41 arquivos de teste, um para cada usuário, que simulam o deslocamento dos mesmos (um histórico de localização). Estes arquivos de teste serviram para testar a eficácia tanto das regras que buscam inferir a ocupação do usuário quanto da rede bayesiana que armazena o perfil probabilístico do usuário. Os dados que foram tabulados também auxiliaram a medir quão próximo cada perfil gerado pelas técnicas estava do perfil real.

O objetivo dessa entrevista inicial era fazer um pré-teste tanto para validar o questionário aplicado quanto para avaliar equívocos existentes nas regras que inferem a ocupação ou na rede bayesiana. Após serem realizados os devidos ajustes, inclusive no formulário de entrevistas (ver Apêndice B), foi feita uma nova rodada de entrevistas (teste final). Desta vez, 202 pessoas foram abordadas, em pontos de ônibus distintos dos que foram utilizados na primeira rodada (ver Figura 5.1).

Os locais marcados em amarelo indicam os pontos de ônibus utilizados no pré-teste e os marcados em azul os do teste final. Nesta segunda etapa, a quantidade de pontos de ônibus utilizados foi maior e mais diversificada, como os próximos a *shopping centers* e faculdades. Também foi incluído no trajeto o principal terminal de ônibus que a cidade possui atualmente, localizado na parte alta. Buscou-se selecionar, de maneira aleatória, os pontos que possuem um grande fluxo de pessoas, distribuídos tanto na parte alta quanto na parte baixa da cidade. Após a realização das entrevistas de rua, seguiu-se o mesmo padrão citado anteriormente, em que os dados organizados deram origem a 202 arquivos de teste que buscam simular os deslocamentos dos usuários em sua rotina.

Figura 5.1: Mapa que indica os locais de aplicação dos questionários.



Fonte: Adaptado de Google Maps®, 2012.

É importante mencionar que a quantidade de pessoas utilizadas nesse segundo teste não foi aleatória, e sim baseada no serviço online disponibilizado tanto por *Raosoft*¹ quanto por *Creative Research Systems*². Este serviço, que se baseia em uma distribuição normal ou Gaussiana, determina, a partir do tamanho da população-alvo e da margem de erro e do nível de confiança desejados, o tamanho mínimo da amostra para atingir estes níveis. Dada a atual parcela da população que utiliza o sistema de transporte coletivo - aproximadamente 350 mil usuários, segundo a Superintendência Municipal de Transportes e Trânsito - SMTT³ - estabelecemos uma margem de erro de 6,5% e um nível de confiança de 93,5% para chegarmos a uma amostra de aproximadamente 200 pessoas. Estes cálculos são feitos supondo uma distribuição uniforme, ou seja, que as pessoas entrevistadas conseguem representar os demais usuários não entrevistados.

Outro ponto que deve ser mencionado é que no pré-teste foram gerados históricos que continham três dias, enquanto no teste final eles continham sete dias.

¹<http://www.raosoft.com/samplesize.html>

²<http://www.surveysystem.com/sscalc.htm>

³<http://goo.gl/ZNqQZ>

Para estes dados, não foi avaliado o *cold start*, ou seja, qual o número de acessos ou tempo de acesso regular mínimo para obter resultados confiáveis e satisfatórios na recomendação. Isto foi deixado a cargo dos trabalhos futuros.

5.2 Estruturação segundo o método GQM

Para organizar nosso estudo de caso, escolhemos o método GQM, que foi detalhado na Seção 2.5. Assim, seguem abaixo os **Goals**, **Questions** e **Metrics** que precisamos definir.

5.2.1 Goal

De acordo com o que foi dito anteriormente, o primeiro passo é apresentar o *Goal* ou conjunto de *Goal's*. No presente trabalho, temos os seguintes:

- G1: Medir a eficácia da inferência da ocupação dos usuários, dado o universo estudante e assalariado, a partir dos dados de deslocamento gerados pelas respostas fornecidas nos questionários;

Objeto de estudo: O objeto de estudo são as regras criadas para identificar a ocupação do usuário baseado nos deslocamentos do mesmo.

Propósito: Verificar se as regras foram construídas de forma adequada.

Foco da Qualidade: Tendo a ocupação um papel significativo na qualidade do perfil que será gerado, é importante avaliar se esta etapa está ocorrendo de forma adequada.

Perspectiva: A perspectiva é do ponto de vista do usuário.

- G2: Dado um anunciante qualquer, aferir a similaridade entre o perfil dos usuários com relação ao público-alvo do anunciante, dados os três cenários/estratégias abaixo, em que X representa uma quantidade qualquer, mas a mesma nos três:

1. Cenário 1: Escolha aleatória de X usuários para uma determinada amostra, utilizando como perfil o próprio perfil informado;
2. Cenário 2: Escolher os X primeiros usuários dada a criação do perfil dos mesmos utilizando puramente os dados georreferenciados;
3. Cenário 3: Escolher os X primeiros usuários dada a criação do perfil dos mesmos utilizando os dados georreferenciados em conjunto com a ocupação.

Objeto de estudo: O objeto de estudo é a compatibilização do perfil em cada uma das situações.

Propósito: Verificar em qual cenário temos uma maior compatibilidade entre o público-alvo do anunciante e o perfil obtido do usuário, possibilitando uma melhor recomendação e um retorno de investimento (ROI) mais atraente. Na tentativa de

medir a contribuição e avaliar o trabalho, foram criados estes três cenários como uma estratégia à comparação com os trabalhos relacionados. Devido à forma como estes trabalhos foram publicados, houve dificuldades de reproduzir as propostas/serviços que existem na literatura.

Foco da Qualidade: Dado que a proposta do trabalho é criar um perfil probabilístico/automatizado baseado em dados georreferenciados em conjunto com a ocupação, espera-se que este perfil seja o que está mais alinhado ao perfil real do usuário.

Perspectiva: A perspectiva é do ponto de vista do usuário.

5.2.2 Questions

Dados os objetivos, nesta seção vamos refiná-los, definindo algumas questões para cada um deles: Para o G1, temos:

- Q1: Qual a proporção de inferências corretas, quanto à ocupação das pessoas? Para o G2, temos:
- Q2: Qual a vantagem que obtemos ao utilizar além dos dados georreferenciados a ocupação, para a inferência dos perfis?

Para Q2, podemos medir esta vantagem comparando as métricas descritas a seguir, entre os três cenários, e analisando qual se saiu melhor, ou seja, qual dos cenários retornou os perfis mais próximos ao público-alvo dos anunciantes.

5.2.3 Metrics

Com os objetivos refinados, precisamos estabelecer um conjunto de métricas para cada um deles.

Para Q1, temos:

- M1:
$$\frac{\text{quantidade de ocupações inferidas corretamente}}{\text{quantidade de usuários}}$$

Para Q2, temos:

- M2: Média aritmética
- M3: Mediana
- M4: Desvio padrão
- M5: Erro quadrático
- M6: Erro máximo absoluto

Para a métrica definida para Q1, temos algo bastante simples: uma proporção de acertos

com relação ao total, resultando em uma porcentagem. Já para as métricas definidas para Q2, temos uma série de medidas estatísticas que isoladamente são simplórias e podem não permitir uma análise adequada dos dados, mas que em conjunto conseguem dar indícios de como os dados estão se comportando. Para cada um dos cenários descritos anteriormente, cada um dos usuários tem seu perfil representado por um vetor e o mesmo é comparado ao público-alvo do anunciante (também representado por um vetor) através de uma métrica de similaridade conhecida por Cosseno de Salton (para mais detalhes sobre os vetores de perfis, ver Capítulo 6). Adiante, a similaridade de cada um dos usuários formam uma lista de similaridades, do qual será extraído a média aritmética, a mediana, o desvio padrão, o erro quadrático e o erro máximo absoluto. Dessa forma, quanto mais próximo um cenário estiver de M2, M3, M4 e M5, melhor será o desempenho dele.

5.3 Definição da avaliação dos resultados

Descritas as métricas para cada *Goal*, vamos definir como será a forma de avaliação para cada uma delas. No caso de M1, ao obtermos uma proporção de acertos maior que 51%, iremos considerar que as regras estão minimamente satisfatórias, visto que temos uma taxa de acertos acima da metade da quantidade de usuários. Com relação à M2, quando a estratégia que utiliza os dados georreferenciados em conjunto com a ocupação (Cenário 3) resultar em um valor maior que a abordagem puramente aleatória (Cenário 1) e também maior que a abordagem que utiliza somente os dados georreferenciados (Cenário 2), iremos considerá-la inicialmente satisfatória, dado que a média é superior as dos demais cenários. O mesmo caso se aplica a M3. Quanto maior a média e a mediana, maior a compatibilidade entre o perfil do usuário e o público-alvo.

Já para M4, M5 e M6, quando o **Cenário 3** resultar em um valor menor que o **Cenário 1** e menor que o **Cenário 2**, é que iremos considerar inicialmente satisfatório, dado que quanto menor o desvio padrão e o erro, melhor será considerado o cenário. O Capítulo 6 abordará, mais detalhadamente, os resultados obtidos.

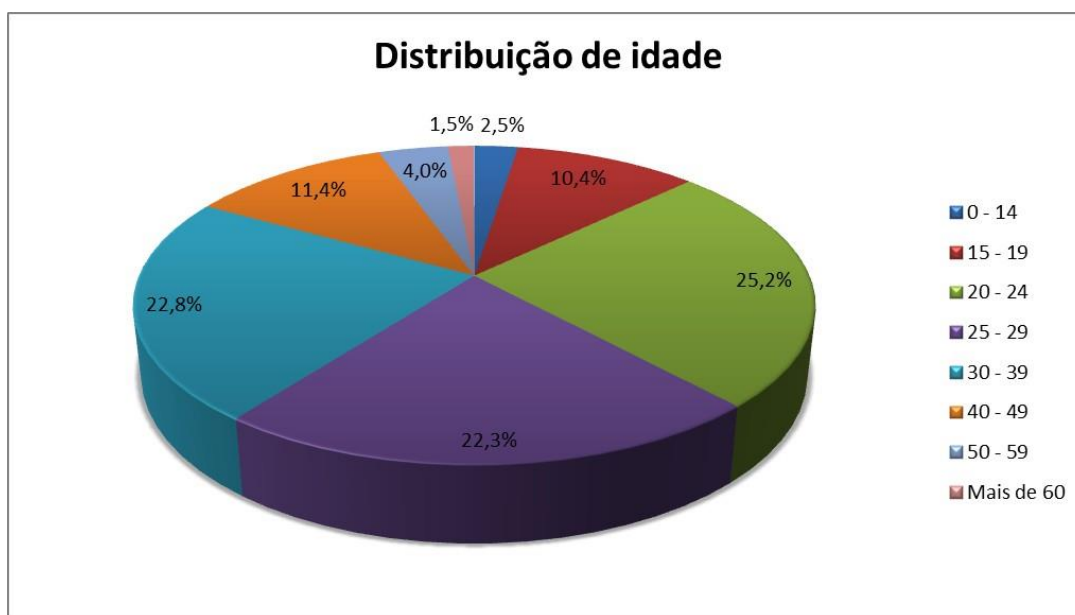
6 Análise dos resultados

No Capítulo 5, foram estabelecidos alguns objetivos a serem alcançados, e os mesmos terminaram por serem definidos através de métricas. Para dar maiores detalhes sobre as métricas apresentadas, temos as seções a seguir.

6.1 Perfil dos entrevistados

Retomando o que foi dito, na realização das entrevistas procurou-se abordar usuários das mais diversas idades, classes sociais e ocupações. Tomando como exemplo a ocupação, a base de usuários da primeira rodada de entrevistas possui 5 estudantes, 22 assalariados, 9 assalariados e estudantes, 2 aposentados, 2 autônomos e 1 desempregado. Já para a segunda rodada de entrevistas, podemos entrar em mais detalhes através dos gráficos a seguir, que descrevem a distribuição de idade (Gráfico 6.1), gênero (Gráfico 6.2), escolaridade (Gráfico 6.3), renda (Gráfico 6.4), ocupação (Gráfico 6.5) e frequência média de utilização dos coletivos (Gráfico 6.6) para os mais de 200 usuários abordados.

Figura 6.1: Gráfico que mostra a distribuição de idade dos entrevistados.



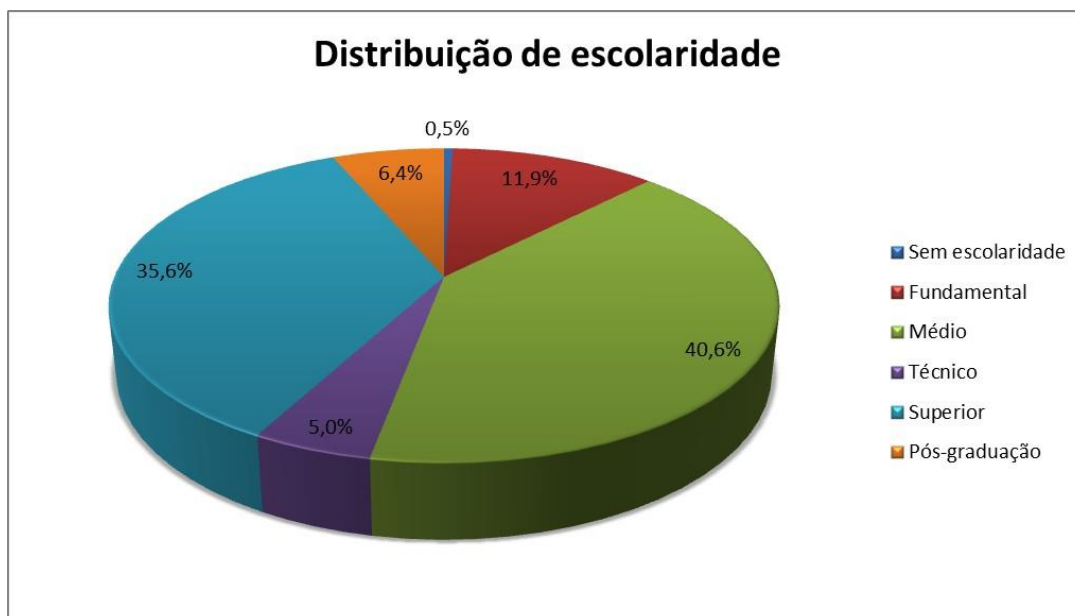
Fonte: Autor, 2012.

Figura 6.2: Gráfico que ilustra a configuração dos entrevistados, por gênero.



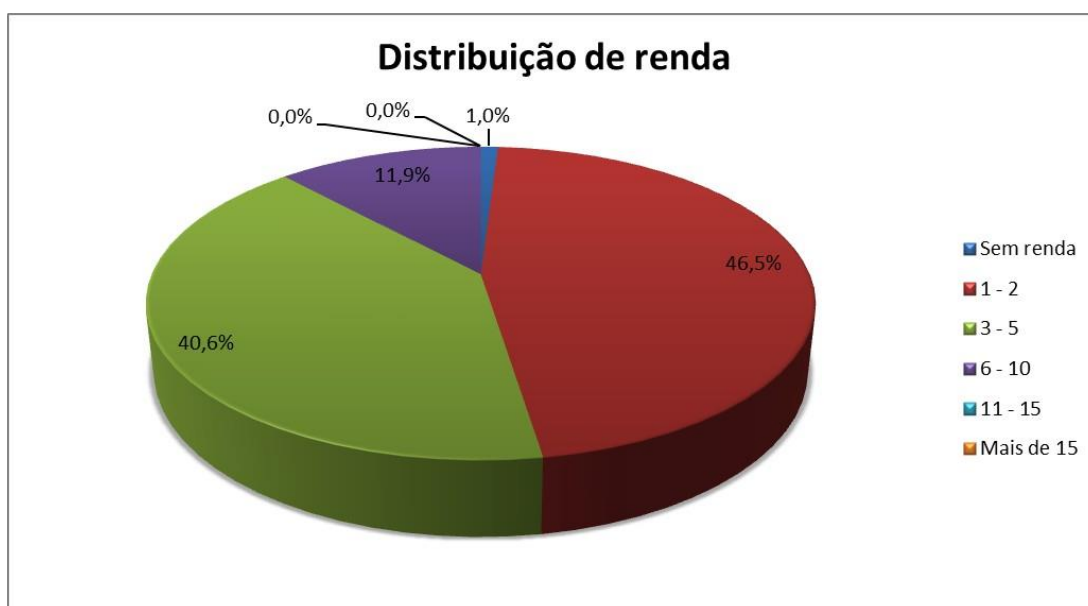
Fonte: Autor, 2012.

Figura 6.3: Gráfico que exibe a distribuição de escolaridade entre os entrevistados.



Fonte: Autor, 2012.

Figura 6.4: Gráfico que retrata a distribuição de renda dos entrevistados, em salários-mínimos.



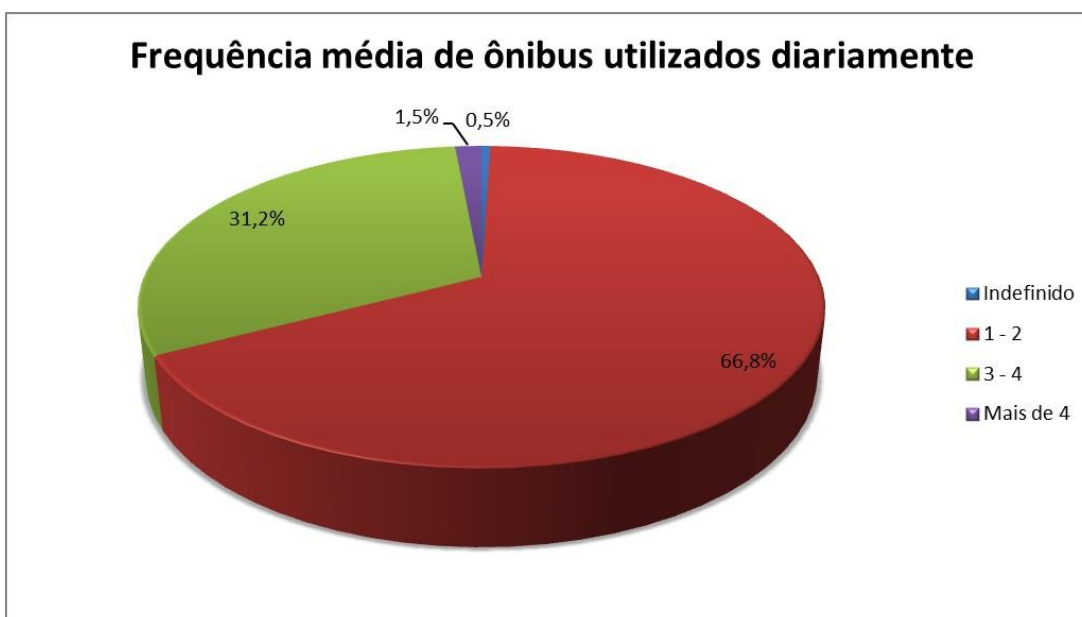
Fonte: Autor, 2012.

Figura 6.5: Gráfico que aborda a distribuição de ocupação dos entrevistados.



Fonte: Autor, 2012.

Figura 6.6: Gráfico que mostra a frequência média de utilização dos ônibus pelos entrevistados.



Fonte: Autor, 2012.

6.2 Público-alvo dos anunciantes

Para a execução do nosso estudo de caso, decidimos escolher cinco potenciais empresas como anunciantes. Dessa forma, nossa base de anunciantes é formada por um centro de estética, uma autoescola, uma farmácia, uma concessionária de veículos e uma empresa de tecnologia que oferece serviços para usuários de *smartphones*. Buscou-se escolher anunciantes distintos e aleatórios para mostrar que a proposta não se aplica somente a um contexto ou determinado público-alvo, mas que ela pode se adequar a vários contextos. Além disso, se tivéssemos utilizado anunciantes semelhantes, isto poderia viciar os resultados do trabalho, restringindo a avaliação do mesmo. Para avaliar isto corretamente, talvez seja adequado adicionar aos trabalhos futuros como o nosso estudo de caso se comportaria com anunciantes semelhantes.

Para cada um destes, produzimos, também através de entrevistas com os mesmos, os respectivos cinco perfis de público-alvo que podem ser vistos abaixo:

Anunciante 1: Concessionária de veículos

Faixa de idade: 25 a 55.

Gênero: Ambos.

Escolaridade: Ensino Médio a Superior.

Renda: A partir de 4 salários-mínimos.

Ocupação: Assalariado.

Anunciante 2: Farmácia

Faixa de idade: A partir de 15 anos.

Gênero: Ambos.

Escolaridade: Todas.

Renda: De 1 a 10 salários-mínimos.

Ocupação: Assalariado e estudante.

Anunciante 3: Centro de estética

Faixa de idade: 20 a 49.

Gênero: Ambos.

Escolaridade: Ensino Superior.

Renda: A partir de 6 salários-mínimos.

Ocupação: Assalariado e estudante.

Anunciante 4: Auto-escola

Faixa de idade: 18 a 25.

Gênero: Ambos.

Escolaridade: A partir do Ensino Médio/Técnico.

Renda: A partir de 1 salário-mínimo.

Ocupação: Assalariado e estudante.

Anunciante 5: Empresa de tecnologia

Faixa de idade: 13 a 25.

Gênero: Ambos.

Escolaridade: Todas.

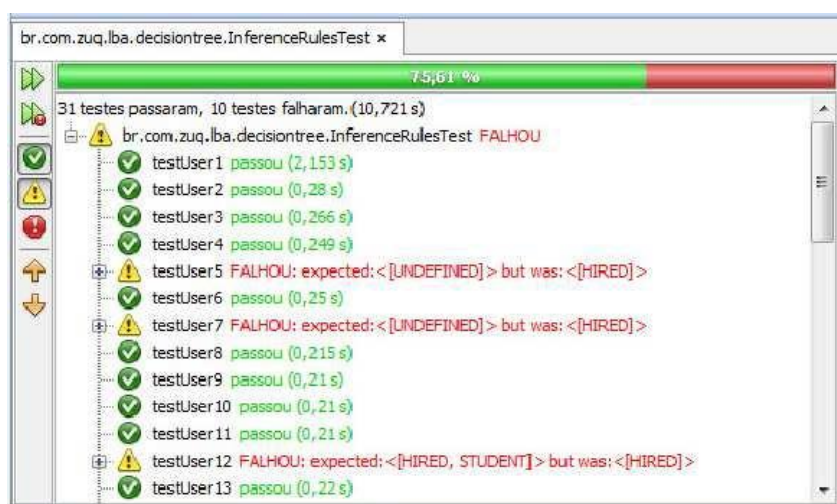
Renda: De 2 a 8 salários-mínimos.

Ocupação: Assalariado e estudante.

6.3 Métrica para G1

Anteriormente, para medir a qualidade da inferência da ocupação dos usuários, foi estabelecido que $M1 = \frac{\text{quantidade de ocupações inferidas corretamente}}{\text{quantidade de usuários}}$. Para fazer esta medição, foi necessário entrevistar usuários do transporte público coletivo, obtendo dados sobre a rotina de uso dos ônibus, dados demográficos e ocupação dos mesmos. De posse destes dados, foram criados arquivos que simulassem um histórico de localização dos usuários (ver Capítulo 5). Em seguida, buscou-se inferir a ocupação de cada um deles através da **Base de Regras** e comparou-se com a informação correta, informada no questionário. O resultado dos usuários entrevistados no pré-teste pode ser visto na Figura 6.7. Nesta figura, “passou” significa que a base de regras inferiu corretamente a ocupação do usuário, enquanto “FALHOU” indica o contrário ou que não foi possível inferir a ocupação.

Figura 6.7: Resultado dos testes das regras da Base de Regras.



Fonte: Autor, 2012.

Como dito anteriormente, a nossa solução somente abrange assalariados e estudantes, o que implica que teremos no máximo 87,8% de acerto, se considerarmos todos os usuários entrevistados. Para este conjunto, atingimos o limiar de 75,61%. Porém, se considerarmos somente o universo de estudantes e assalariados, a taxa de acerto sobe para 86,1%. Devido ao fato de algumas pessoas dependerem parcialmente do transporte coletivo, em ambas as ocupações, muitas vezes pegando carona com amigos ou família, a nossa precisão não pode ser maior. Em outros casos, os usuários são estudantes e assalariados mas possuem o comportamento somente de assalariados, já que estudam e trabalham ou no mesmo local ou próximos um do outro, o que exclui a necessidade do coletivo para locomoção. Com relação ao segundo grupo de usuários entrevistados, chegamos ao limiar de acerto de 76,2%, em um total de 94,5%, já que temos 11 usuários que não são nem assalariados nem estudantes. Restringindo os mesmos ao universo descrito anteriormente, a taxa de acertos sobe para 80,6%. Para esse grupo de usuários, encontramos um novo fator que impactou na taxa de acertos: alguns deles utilizam o transporte alternativo em alguns momentos da sua locomoção diária. Observando os resultados apresentados, vemos que eles estão de acordo com o limiar mínimo estabelecido na Seção 5.3, mostrando que nossa base de regras está potencialmente satisfatória.

6.4 Métricas para G2

Em G2, desejamos aferir a similaridade entre o perfil dos usuários e o público-alvo do anunciante, nos três cenários apresentados, para um anunciante qualquer. Para perfazer o conjunto de métricas para o **G2**, após a entrevista aos usuários de transporte coletivo, descrita na seção anterior, guardou-se a informação dos bairros onde os mesmos moravam, estudavam e/ou trabalhavam para inferir seus perfis probabilísticos. Com estas informações, três conjuntos de perfis (cenários) puderam ser criados:

Cenário aleatório: dada uma amostra de usuários, para este cenário ficou estabelecido que seriam escolhidos Z usuários aleatórios, em que Z é definido como $\frac{1}{4}$ do tamanho da amostra. Escolhidos os Z usuários, cada um deles é comparado ao perfil de público-alvo do anunciante, formando um conjunto de valores de similaridade entre anunciante-usuário. Esta tarefa é realizada 100 vezes, para fins de evitar/remover algum vício na geração aleatória, e o cenário escolhido é o que apresenta a melhor média e, em caso de empate, o menor desvio padrão, ou seja, sempre é utilizado o melhor cenário aleatório.

Cenário georreferenciados: nesta abordagem, para cada usuário, utilizaram-se os dados georreferenciados do local de residência para estimar todos os atributos de seu perfil, composto por idade, gênero, escolaridade e renda. Obtidos estes dados, os mesmos foram passados ao elemento **Módulo da Rede Bayesiana**, que guarda e infere um perfil probabilístico. Ao fim, no elemento **Módulo de Recomendação**, os usuários são ordenados

pela similaridade, de modo decrescente, e os P primeiros usuários são usados para formar este cenário. Do mesmo modo do cenário aleatório, P é definido como $\frac{1}{4}$ do tamanho da amostra.

Cenário georreferenciados com ocupação: para este perfil, temos a mesma mecânica do perfil descrito acima, adicionado somente o atributo ocupação, com as probabilidades a priori descritas no Capítulo 5. Neste caso, após inferirmos um perfil probabilístico, a ocupação é inserida, obtida a partir da Base de Regras, como uma evidência e a rede é recompilada.

O fato de escolhermos P e Z como $\frac{1}{4}$ do tamanho da amostra justifica-se por tentarmos tratar com valores mais exatos, já que nosso estudo de caso possui 202 pessoas. Da mesma forma poderíamos também ter escolhido $\frac{1}{2}$, mas com 50 pessoas os conjuntos de usuários podem ser mais diversificados, já que temos a impressão que escolher 50 pessoas entre 200 tem a possibilidade de gerar conjuntos mais diversificados do que escolher 100 entre 200 pessoas. O resultado, do que foi descrito acima, pode ser visto nas tabelas a seguir, somente para a segunda etapa (teste final), já que a primeira rodada de entrevistas (pré-teste) teve o objetivo de ajuste. Para resultados mais detalhados, ver Apêndice C.

Tabela 6.1: Valores do teste final para o público-alvo da concessionária de veículos

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,518	3,023	4,178
Mediana	2,689	3,009	4,168
Desvio Padrão	0,513	0,161	0,210
Erro quadrático	6,588	4,016	0,733
Erro máximo absoluto	4,293	2,251	1,110

Fonte: Autor, 2012.

Tabela 6.2: Valores do teste final para o público-alvo da farmácia

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,729	3,321	4,169
Mediana	2,646	3,318	4,167
Desvio Padrão	0,147	0,070	0,075
Erro quadrático	5,291	2,882	0,709
Erro máximo absoluto	2,631	1,802	0,929

Fonte: Autor, 2012.

Tabela 6.3: Valores do teste final para o público-alvo do centro de estética

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,344	2,784	3,629
Mediana	2,571	2,768	3,575
Desvio Padrão	0,603	0,268	0,261
Erro quadrático	7,587	5,082	1,984
Erro máximo absoluto	3,884	2,614	1,672

Fonte: Autor, 2012.

Tabela 6.4: Valores do teste final para o público-alvo da auto escola

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,731	3,108	3,773
Mediana	2,640	3,082	3,721
Desvio Padrão	0,523	0,062	0,124
Erro quadrático	5,531	3,655	1,551
Erro máximo absoluto	3,432	1,956	1,369

Fonte: Autor, 2012.

Tabela 6.5: Valores do teste final para o público-alvo da empresa de Tecnologia da Informação

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,729	3,151	3,792
Mediana	2,805	3,152	3,758
Desvio Padrão	0,457	0,073	0,129
Erro quadrático	5,501	3,494	1,506
Erro máximo absoluto	3,576	1,945	1,377

Fonte: Autor, 2012.

6.5 Discussão dos resultados

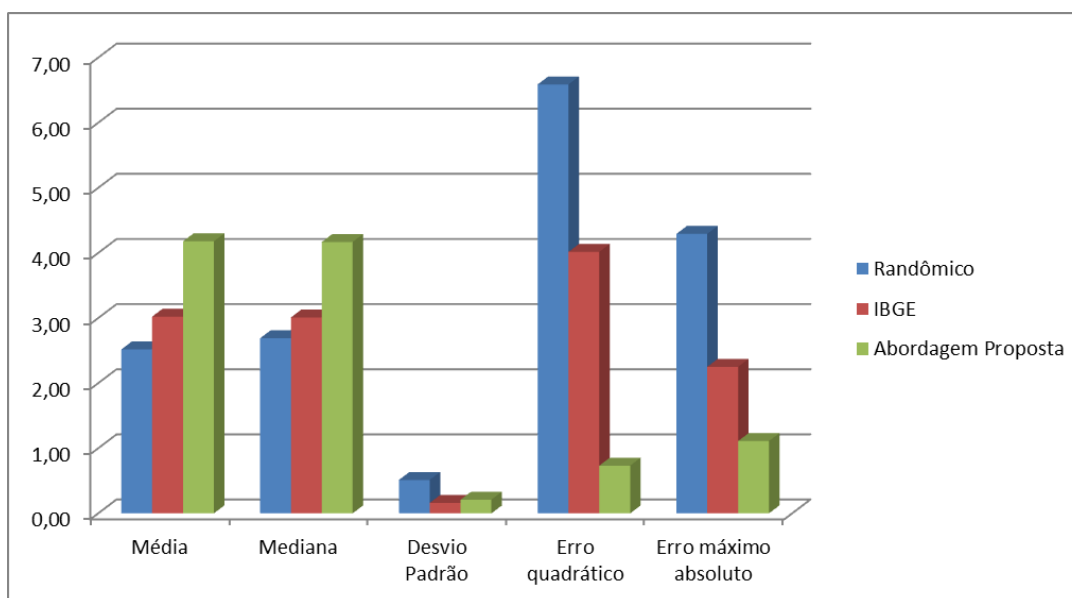
De forma geral, pelos resultados mostrados nas Tabelas 6.1, 6.2, 6.3, 6.4 e 6.5 e nas Figuras 6.8, 6.9, 6.10, 6.11 e 6.12, percebe-se que a abordagem proposta (dados georreferenciados juntamente com ocupação) apresenta médias e medianas um pouco melhores e erros quadráticos e absolutos menores, quando comparado aos demais cenários. Dado que foi utilizado o perfil de uma população de dez anos atrás (Censo IBGE 2000) e com anunciantes diversificados e aleatórios, os resultados apresentados estão minimamente satisfatórios. Também podemos perceber que para perfis mais gerais (e.g. concessionária e farmácia), a abordagem proposta possui médias e medianas mais altas, assim como o erro quadrático e o erro máximo absoluto mais baixos. Já no caso de um perfil mais específico, como o centro de estética, que apresenta o perfil mais restrito quanto a idade, educação e renda, a nossa abordagem apresenta as menores média e mediana e o maior desvio padrão, quando comparamos a mesma técnica com os demais anunciantes, pois neste caso o cenário puramente

georreferenciado se aproximou mais.

Acreditamos que essa pequena diminuição no desempenho da técnica proposta, neste último caso, tenha ocorrido em virtude dos usuários entrevistados estarem fora do perfil desejado, principalmente no tocante a educação e renda, e não puramente pelo fato do perfil ser restrito. Ou seja, se trabalharmos com um perfil-alvo restrito, mas que seja próximo ao perfil encontrado no dia-a-dia nos coletivos, provavelmente obteremos resultados satisfatórios. Dessa forma, para obter bons resultados, no contexto do estudo de caso, é importante buscar anunciantes cujo perfis desejados sejam compatíveis com o perfil dos usuários de transporte público. Como mostrado anteriormente, este perfil pode ser consultado em [Carvalho & Santos \(2006\)](#).

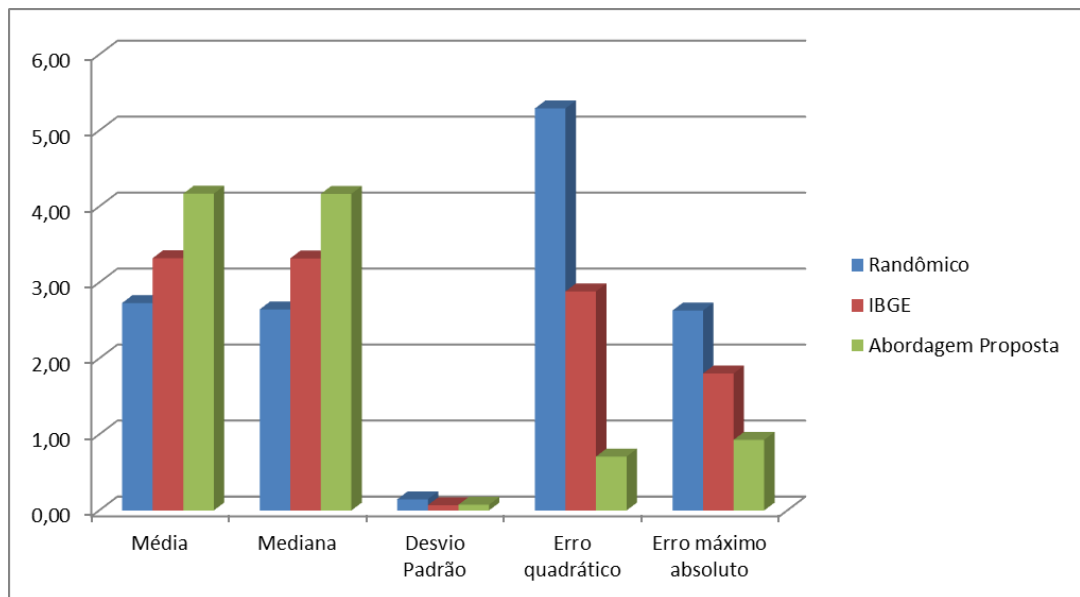
Pelos resultados apresentados, temos um indicativo de que é possível automatizar a personalização de anúncios/ofertas e que o método proposto por este trabalho pode apresentar bons resultados, pelo menos no contexto de usuários do transporte público. Para isto, ainda precisamos analisar vários aspectos, entre eles quais usuários estariam realmente interessados nos anúncios enviados e estabelecer uma porcentagem mínima de aceitação das pessoas, para definir se os resultados estão ou não satisfatórios.

Figura 6.8: Gráfico do teste final para a concessionária de veículos nos três cenários.



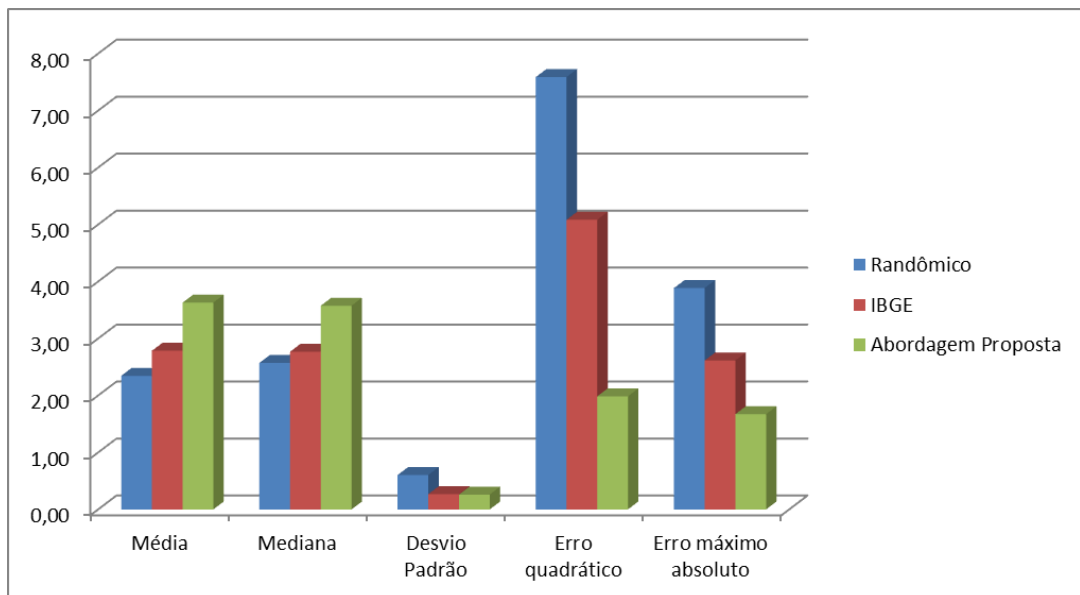
Fonte: Autor, 2012.

Figura 6.9: Gráfico do teste final para a farmácia nos três cenários.



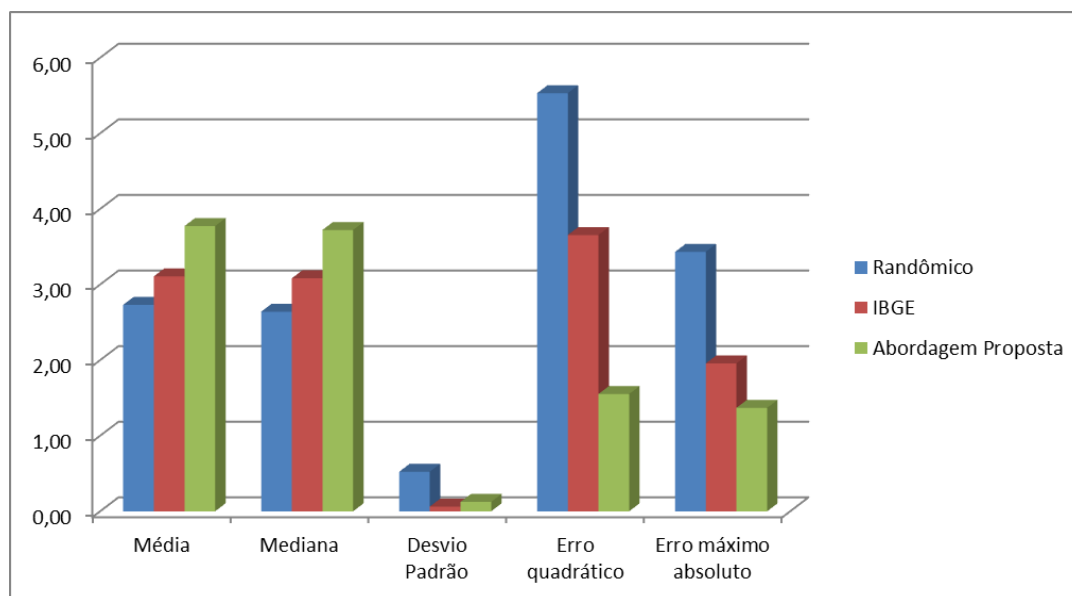
Fonte: Autor, 2012.

Figura 6.10: Gráfico do teste final para o centro de estética nos três cenários.



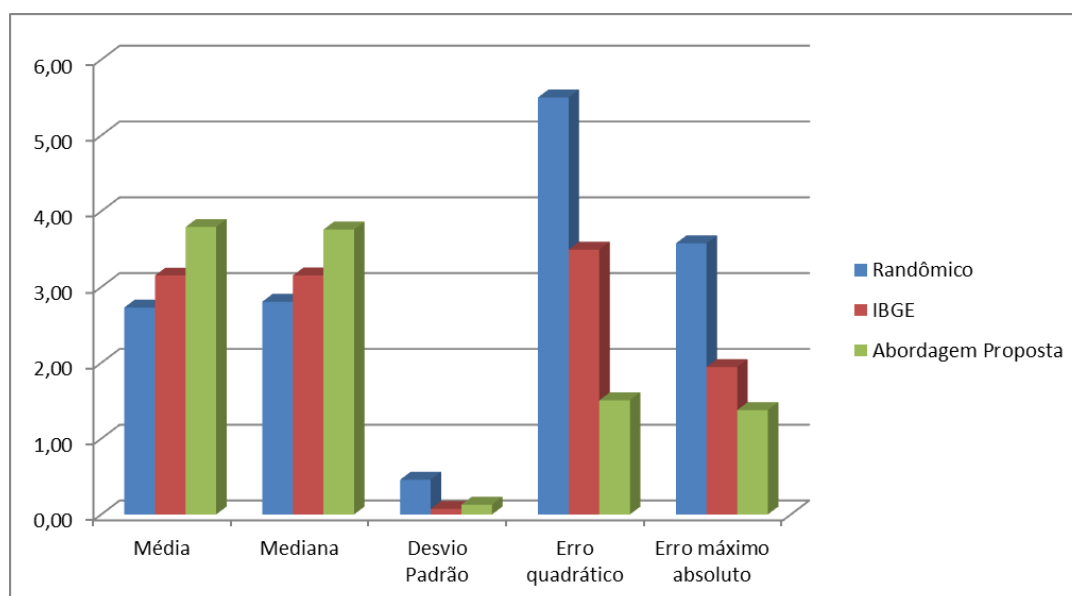
Fonte: Autor, 2012.

Figura 6.11: Gráfico do teste final para a autoescola nos três cenários.



Fonte: Autor, 2012.

Figura 6.12: Gráfico do teste final para a empresa de Tecnologia da Informação nos três cenários.



Fonte: Autor, 2012.

7 Conclusão e Trabalhos Futuros

Location-based service é uma área bastante abrangente e promissora, e com grandes indícios de crescimento, principalmente no Brasil, devido aos altos investimentos feitos em publicidade móvel e também para acompanhar a crescente utilização de dispositivos móveis, principalmente *smartphones*. Muitos aspectos ainda devem ser discutidos até sua popularização, principalmente os relacionados à forma de abordagem em *location-based advertising*.

No Capítulo 2, apresentamos os principais conceitos utilizados no desenvolvimento do trabalho, como *Location-based advertising*, Redes Bayesianas, Árvores de Decisão e Métricas de Similaridade. Além de apresentar conceitos, foram dados alguns exemplos que permitissem uma leitura contextualizada do trabalho.

O Capítulo 4 detalha as principais contribuições que esperamos fornecer tanto para a academia quanto para a indústria. Para facilitar o entendimento, o capítulo ainda foi enriquecido com um exemplo bem simples, mas que permite visualizar uma aplicação prática. O Capítulo 6 mostra o estudo de caso, que apresentou resultados favoráveis à nossa abordagem de criação e refinamento do perfil, quando comparado aos demais cenários, no contexto do transporte público, mostrando bons indícios que estamos na direção certa para fornecer uma boa colaboração à área.

As principais dificuldades foram encontrar dados de localização real, já que os dados do estudo de caso foram aproximados com base nos bairros de residência e/ou estudo e trabalho dos usuários. Outro ponto que poderia ser melhorado é utilizar dados georreferenciados mais atuais, como por exemplo, do Censo 2010, mas que ainda não estão disponíveis.

Atualmente, o serviço proposto está em fase de testes, utilizando como dados georreferenciados os do Censo 2000. Para o estudo de caso, foram feitas duas etapas de entrevistas: a primeira com 41 pessoas e a segunda com 202 pessoas, sempre com usuários que utilizam o transporte público para trabalhar e/ou estudar. Além disso, a base de regras para inferir ocupação aborda, atualmente, somente estudantes e/ou assalariados.

7.1 Principais contribuições da pesquisa

O levantamento do estado da arte nos mostrou que ainda existem algumas limitações nas abordagens existentes para serviços personalizáveis e baseados/sensíveis à localização, principalmente na área de publicidade. Além disso, vários aspectos precisam ser evoluídos e explorados para termos serviços mais atraentes e customizados ao interesse das pessoas. No presente trabalho, a utilização de dados georreferenciados juntamente com históricos de localização e a inferência da ocupação dos usuários buscou mudar esta realidade.

O desenvolvimento desta pesquisa e a avaliação dos resultados buscou contribuir tanto para a academia quanto para a indústria. No caso da academia, buscou-se mostrar que é possível

automatizar a personalização de anúncios, através dos atributos idade, gênero, escolaridade, renda e ocupação (obtidos a partir de base de dados georreferenciadas) e também via históricos de localização das pessoas, estratégia não encontrada na literatura, até o momento. Já para a indústria, é necessário avançar na fase de testes e colocar o sistema resultante da proposta em um ambiente de produção, permitindo avaliá-lo em ambiente real.

7.2 Trabalhos Futuros

Como trabalhos futuros, propomos as contribuições abaixo:

1. Integrar a Base de Regras com ferramentas de classificação, utilizando, por exemplo, classificação com base em um conjunto de treinamento a priori. Desta forma, teríamos um mecanismo de aquisição automática de conhecimento;
2. Adicionar mais um **Goal** à nossa proposta, descrito por: verificar o potencial de retorno de investimento (ROI) dos anúncios, com base em respostas de questionários. Isto levaria a seguinte pergunta: qual a proporção de usuários que de fato estariam interessados em anúncios do tipo recomendado? Desta forma, poderíamos avaliar a compatibilidade dos clientes recomendados a um determinado anúncio, comparando caso isso fosse feito baseado em seu perfil real e o seu perfil gerado pelo sistema;
3. Avaliar de modo preciso o *cold start* para a proposta, ou seja, definir um período mínimo de dias para que o histórico de localização impacte em resultados confiáveis na recomendação dos anúncios;
4. Atualizar a base de dados georreferenciadas, utilizando os dados do Censo 2010;
5. Validar o serviço com base em dados reais de localização;
6. Evoluir a abordagem proposta para integrar elementos da Ciência de Contexto, permitindo ao usuário a obtenção de conteúdo mais próximo ao seu dia-a-dia. Por exemplo, para os usuários que dependem do transporte público, buscar conteúdo mais próximo ao itinerário realizado pelos ônibus utilizados rotineiramente;
7. Planejar e executar uma avaliação quantitativa tanto para a inferência da ocupação quanto para a similaridade entre perfil do público-alvo e perfil dos usuários;
8. Executar outros testes relacionados ao estudo de caso, desta vez com anunciantes similares;
9. Publicar artigos em eventos e congressos da área, principalmente em conferências com um bom *Qualis*, visando divulgar o trabalho desenvolvido para a comunidade científica.

Referências

- Aalto, L., Göthlin, N., Korhonen, J. & Ojala, T. (2004), Bluetooth and wap push based location-aware mobile advertising system, *in* 'Proceedings of the 2nd international conference on Mobile systems, applications, and services', MobiSys '04, ACM, New York, NY, USA, pp. 49–58 DOI <http://doi.acm.org/10.1145/990064.990073>. Disponível em <<http://doi.acm.org/10.1145/990064.990073>> Acessado em: 20 jun.2012.
- Adnews (2012), 'Investimento em publicidade móvel deve atingir us\$ 6,4 bi em 2012'. Disponível em <<http://goo.gl/cXVku>> Acessado em: 20 jun.2012.
- Alves, V. (2012), 'Investimentos em anúncios online devem crescer quase 40% neste ano'. Disponível em <<http://goo.gl/jVYLR>> Acessado em: 20 jun.2012.
- Azevedo, T. (2012), 'Investimento global em publicidade cresce 2,4% no segundo trimestre'. Disponível em <<http://goo.gl/8DZbY>> Acessado em: 20 jun.2012.
- Barbosa, M. (2012), 'Investimento em publicidade cresce 16%'. Disponível em <<http://goo.gl/xurHb>> Acessado em: 20 jun.2012.
- Basili, V., Caldiera, G. & Rombach, D. H. (1994), The goal question metric approach, *in* J. Marciniak, ed., 'Encyclopedia of Software Engineering', Wiley.
- Basili, V. R. & Weiss, D. M. (1984), 'A methodology for collecting valid software engineering data', *IEEE Transactions on Software Engineering* **10**(6), 728–738 Disponível em <<http://dblp.uni-trier.de/rec/bibtex/journals/tse/BasiliW84>> Acessado em: 10 set.2012.
- Blackwell, G. (2005), 'Using wi-fi/cellular in p2p positioning', Wi-Fi Planet News. Disponível em <wi-fiplanet.com/news/article.php/3572001> Acessado em: 13 ago.2012.
- Boyce, B., Meadow, C. & Kraft, D. (1995), *Measurement in information science*, Library and information science, Academic Press.
- Bruner, G., I. & Kumar, A. (2007), 'Attitude toward location-based advertising', *Journal of Interactive Advertising*.
- Bulander, R., Decker, M., Schiefer, G. & Kolmel, B. (2005), Comparison of different approaches for mobile advertising, *in* 'Proceedings of the Second IEEE International Workshop on Mobile Commerce and Services', IEEE Computer Society, Washington, DC, USA, pp. 174–182 DOI [10.1109/WMCS.2005.8](http://dx.doi.org/10.1109/WMCS.2005.8). Disponível em <<http://dl.acm.org/citation.cfm?id=1112245.1112728>> Acessado em: 19 mai.2012.

Carvalho, C. H. R. & Santos, M. B. (2006), Mobilidade da população urbana, Technical report, Associação Nacional das Empresas de Transportes Urbanos.

Chuang, S. W. & Krabuanrat, T. (2007), ‘On the future of location-based advertising’, *SCMS Journal of Indian Management* **4**(1), 10–15 Disponível em <<http://www.ncbi.nlm.nih.gov/pubmed/18497183>> Acessado em: 11 jan.2012.

ComputerWorld (2011), ‘Venda de dispositivos móveis cresce 5,6% no 3^o trimestre’. Disponível em <<http://goo.gl/Ef4zX>> Acessado em: 11 jan.2012.

Dao, T. H., Jeong, S. R. & Ahn, H. (2012), ‘A novel recommendation model of location-based advertising: Context-aware collaborative filtering using ga approach’, *Expert Syst. Appl.* **39**, 3731–3739 DOI <http://dx.doi.org/10.1016/j.eswa.2011.09.070>. ISSN 0957-4174. Disponível em <<http://dx.doi.org/10.1016/j.eswa.2011.09.070>> Acessado em: 20 jun.2012.

Darwiche, A. (2009), *Modeling and Reasoning with Bayesian Networks*, Cambridge University Press.

Egghe, L. (2009), ‘New relations between similarity measures for vectors based on vector norms’, *Journal of the American Society for Information Science and Technology* **60**(2), 232–239 DOI [10.1002/asi.v60:2](http://dx.doi.org/10.1002/asi.v60:2). ISSN 1532-2882. Disponível em <<http://dx.doi.org/10.1002/asi.v60:2>> Acessado em: 14 mar.2012.

Egghe, L. & Michel, C. (2002), ‘Strong similarity measures for ordered sets of documents in information retrieval’, *Information Processing and Management* **38**(6), 823 – 848 DOI [10.1016/S0306-4573\(01\)00051-6](http://dx.doi.org/10.1016/S0306-4573(01)00051-6). ISSN 0306-4573. Disponível em <<http://www.sciencedirect.com/science/article/pii/S0306457301000516>> Acessado em: 18 abr.2012.

Gratton, E. (2002), ‘M-commerce: The notion of consumer consent in receiving location-based advertising’, *Canadian Journal of Law and Technology*.

Grossman, D. & Frieder, O. (1998), *Information Retrieval: Algorithms and Heuristics*, Kluwer international series in engineering and computer science, Kluwer Academic Publishers.

Heckerman, D. (1996), A tutorial on learning with bayesian networks, Technical report, Learning in Graphical Models.

Huang, Y. & Bian, L. (2009), ‘A bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the internet’, *Expert Syst. Appl.* **36**(1), 933–943 DOI [10.1016/j.eswa.2007.10.019](http://dx.doi.org/10.1016/j.eswa.2007.10.019). ISSN 0957-4174. Disponível em <<http://dx.doi.org/10.1016/j.eswa.2007.10.019>> Acessado em: 14 ago.2012.

- IDGNow! (2011), 'Smartphone já chegou às massas no brasil, aponta pesquisa'. Disponível em <<http://goo.gl/YScdE2>> Acessado em: 20 jun.2012.
- Jordan, G. (2012), 'Analistas estimam que mercado brasileiro de celular vai manter ritmo de crescimento'. Disponível em <<http://goo.gl/MY5ZT>> Acessado em: 20 jun.2012.
- Kölmel, B. & Spiros, A. (2002), 'Location based advertising'. Disponível em <<http://goo.gl/wW6gS>> Acessado em: 21 jun.2012.
- Kuo, M.-H., Chen, L.-C. & Liang, C.-W. (2009), 'Building and evaluating a location-based service recommendation system with a preference adjustment mechanism', *Expert Syst. Appl.* **36**, 3543–3554 DOI [10.1016/j.eswa.2008.02.014](https://doi.org/10.1016/j.eswa.2008.02.014). ISSN 0957-4174. Disponível em <<http://dl.acm.org/citation.cfm?id=1464526.1465299>.> Acessado em: 14 mar.2012.
- Küpper, A. (2005), *Location-based services: fundamentals and operation*, John Wiley. Disponível em <<http://books.google.com.br/books?id=4FcZAQAIAAJ>> Acessado em: 03 mai.2012.
- Liapis, D., Vassilaras, S. & Yovanof, G. (2008), Implementing a low-cost, personalized and location based service for delivering advertisements to mobile users, in 'Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on', pp. 133 –137 DOI [10.1109/ISWPC.2008.4556182](https://doi.org/10.1109/ISWPC.2008.4556182).
- Losee, R. (1998), *Text retrieval and filtering: analytic models of performance*, The Kluwer international series on information retrieval, Kluwer Academic Publishers.
- Manning, C. D., Raghavan, P. & Schtze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA.
- Marketer, M. (2009), 'Classic guide to mobile advertising'. Disponível em <<http://goo.gl/P471X>> Acessado em: 26 dez.2011.
- Mazaheri, A., Rafiee, N. & Khadivi, P. (2010), Location based targeted advertising using bayesian network and fuzzy topsis, in 'Telecommunications (IST), 2010 5th International Symposium on', pp. 645 –650 DOI [10.1109/ISTEL.2010.5734103](https://doi.org/10.1109/ISTEL.2010.5734103).
- Nilsson, N. J. (1998), 'Introduction to machine learning: An early draft of a proposed textbook'.
- Panayiotou, C. & Samaras, G. (2006), Mobile user personalization with dynamic profiles: Time and activity, in R. Meersman, Z. Tari & P. Herrero, eds, 'On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops', Vol. 4278 of *Lecture Notes in Computer*

- Science*, Springer Berlin / Heidelberg, pp. 1295–1304. Disponível em <http://dx.doi.org/10.1007/11915072_33> Acessado em: 15 ago.2011.
- Rashid, O., Coulton, P. & Edwards, R. (2008), ‘Providing location based information/advertising for existing mobile phone users’, *Personal Ubiquitous Comput.* **12**, 3–10 DOI <http://dx.doi.org/10.1007/s00779-006-0121-4>. ISSN 1617-4909. Disponível em <<http://dx.doi.org/10.1007/s00779-006-0121-4>> Acessado em: 08 abr.2011.
- Russell, S. J. & Norvig, P. (1995), *Artificial intelligence: a modern approach*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- Salton, G. & McGill, M. J. (1987), *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA.
- Schiller, J. & Voisard, A. (2004), *Location-Based Services*, Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers. Disponível em <<http://books.google.com.br/books?id=wj19b5wVfXAC>> Acessado em: 27 jun.2011.
- Service, I. N. (2010), ‘Vendas mundiais de celulares crescem 14% no 2^o trimestre.’. Disponível em <<http://goo.gl/W87uj>> Acessado em: 27 jul.2011.
- Service, I. N. (2012), ‘Pela 1^a vez na história, entregas de smartphones superam as de pcs’. Disponível em <<http://goo.gl/fkuCa>> Acessado em: 26 jul.2011.
- Shannon, R., Stabeler, M., Quigley, A. & Nixon, P. (2009), Profiling and targeting opportunities in pervasive advertising, in ‘1st Workshop on Pervasive Advertising’, Vol. 1, Pervasive 2009, Nara, Japan.
- Tague-Sutcliffe, J. (1995), *Measuring information: an information services perspective*, Library and information science, Academic Press.
- Van Rijsbergen, C. (1979), *Information retrieval*, Butterworths.
- Van Solingen, R. & Berghout, E. (1999), *The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development*, McGraw-Hill.
- van Solingen, R., Basili, V., Caldiera, G. & Rombach, H. D. (2002), *Goal Question Metric (GQM) Approach*, John Wiley & Sons, Inc.
- Walczak, B. (2005), ‘Profile specific ads make the difference’, Mobile Marketing Association. Disponível em <<http://mmaglobal.com/articles/profile-specific-ads-make-difference>> Acessado em: 07 nov.2011.
- Yaakub, N. & Napiah, M. (2011), Public bus passenger demographic and travel characteristics a study of public bus passenger profile in kota bharu, kelantan, in ‘National Postgraduate Conference (NPC), 2011’, pp. 1–6 DOI [10.1109/NatPC.2011.6136379](https://doi.org/10.1109/NatPC.2011.6136379).

Apêndice A

Questionário para identificação de perfis - Pré-teste

Este questionário visa auxiliar na identificação do perfil de assalariados e estudantes que utilizam ônibus para seus deslocamentos diários. A identificação desses perfis faz parte do desenvolvimento de uma pesquisa de dissertação.

Gênero:

Masculino Feminino

Estado civil:

Solteiro(a) Casado(a) Viúvo(a) Divorciado(a)

Data de nascimento (dd/mm/aaaa):

Qual sua renda familiar?(S.M. significa salário-mínimo(s)):

1 s.m. a 2 s.m. 3 s.m. a 5 s.m. 6 s.m. a 10 s.m. 11 s.m. a 15 s.m.
 Acima de 15 s.m.

Bairro de residência:

Quantas pessoas residem com você?

Qual seu nível de escolaridade?

Básico Fundamental Médio Incompleto Médio Completo
 Técnico Incompleto Técnico Completo Superior Incompleto Superior
Completo Pós-Graduação Incompleto Pós-Graduação Completo

Você está estudando atualmente?

Sim Não

Onde estuda?

Bairro da escola/faculdade/universidade:

Você está trabalhando atualmente?

Sim Não

Qual sua profissão?

Nome do local de trabalho:

Bairro do trabalho:

Quais dias (semanalmente) você costuma utilizar ônibus para se deslocar?

Domingo Segunda-feira Terça-feira Quarta-feira Quinta-feira

Sexta-feira Sábado

Com que frequência, em média, você costuma utilizar ônibus para se deslocar diariamente?

Quais as principais razões para a utilização do ônibus com o meio de deslocamento?

Ir ao trabalho Ir à escola/faculdade/universidade Lazer

Tratamento de saúde Outra opção:

Em sua rotina, qual(is) linha(s) de ônibus você costuma pegar quando sai de casa?

Quais horários você costuma sair de casa quando utiliza ônibus?

Em sua rotina, qual(is) linha(s) de ônibus você costuma pegar ao voltar para casa?

Quais horários você costuma voltar para casa quando utiliza ônibus?

Apêndice B

Questionário para identificação de perfis - Teste final

Este questionário visa auxiliar na identificação do perfil de assalariados e estudantes que utilizam ônibus para seus deslocamentos diários. A identificação desses perfis faz parte do desenvolvimento de uma pesquisa de dissertação.

Contato (e-mail):

Gênero:

Masculino Feminino

Estado civil:

Solteiro(a) Casado(a) Viúvo(a) Divorciado(a)

Data de nascimento (dd/mm/aaaa):

Qual sua renda familiar?(S.M. significa salário-mínimo(s)):

1 s.m. a 2 s.m. 3 s.m. a 5 s.m. 6 s.m. a 10 s.m. 11 s.m. a 15 s.m.
 Acima de 15 s.m.

Bairro de residência:

Qual o total de pessoas em sua residência?

Qual seu nível de escolaridade?

Sem escolaridade Fundamental Médio Incompleto Médio Completo
 Técnico Incompleto Técnico Completo Superior Incompleto Superior Completo

Pós-Graduação Incompleto Pós-Graduação Completo

Você está estudando atualmente?

Sim Não

Onde estuda?

Bairro da escola/faculdade/universidade:

Você está trabalhando atualmente?

Sim Não

Qual sua profissão?

Nome do local de trabalho:

Bairro do trabalho:

Em média, qual a quantidade de ônibus utilizados para se deslocar diariamente?

Quais dias (semanalmente) você costuma utilizar ônibus para se deslocar?

Domingo Segunda-feira Terça-feira Quarta-feira Quinta-feira

Sexta-feira Sábado Indefinido

Quais as principais razões para a utilização do ônibus com o meio de deslocamento?

Ir ao trabalho Ir à escola/faculdade/universidade Lazer

Tratamento de saúde Outras razões

Em sua rotina, qual(is) linha(s) de ônibus você costuma pegar quando sai de casa?

Quais horários você costuma sair de casa quando utiliza ônibus?

Em sua rotina, qual(is) linha(s) de ônibus você costuma pegar ao voltar para casa?

Quais horários você costuma voltar para casa quando utiliza ônibus?

Em sua rotina, você costuma fazer deslocamentos intermediários? (Caso positivo, escolher uma das opções abaixo)

Utiliza ônibus para almoçar em casa

- () Utiliza ônibus para sair do trabalho para local de estudo
- () Utiliza ônibus para sair do local de estudo para o trabalho
- () Outros. Bairro:

Qual(is) linha(s) de ônibus você costuma pegar?

Quais horários você costuma utilizar ônibus para realizar estes deslocamentos?

Apêndice C

Resultados Completos do estudo de caso

Tabela C.1: Valores do pré-teste para o público-alvo da concessionária de veículos

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,738117993652040	2,969894182874960	4,257853855393110
Desvio Padrão	0,147100309208935	0,284699389906643	0,264518395554737
Erro quadrático	5,777262487940080	4,660308885640880	0,681948759314473
Erro máximo absoluto	2,466945015373110	2,466887550741160	1,191562758265420

Fonte: Autor, 2012.

Tabela C.2: Valores do pré-teste para o público-alvo da farmácia

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,80054734123977	3,30135937631671	4,16537135570552
Desvio Padrão	0,16004911292970	0,05131918044753	0,04596113498897
Erro quadrático	5,66947304969825	3,20861140097860	0,77611795245928
Erro máximo absoluto	2,39887287120172	1,73669725785145	0,88244967691118

Fonte: Autor, 2012.

Tabela C.3: Valores do pré-teste para o público-alvo do centro de estética

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,66464213197828	2,73683310379573	3,68011186933911
Desvio Padrão	0,55858416249274	0,34010073111465	0,29378242248475
Erro quadrático	6,54504069142307	5,80669561849901	2,02197997560488
Erro máximo absoluto	3,54406314117563	2,67286172308264	1,68273730356155

Fonte: Autor, 2012.

Tabela C.4: Valores do pré-teste para o público-alvo da auto escola

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	3,09575812177760	3,12519531824988	3,78081556842110
Desvio Padrão	0,45264308636161	0,12742547563664	0,15857943927495
Erro quadrático	4,23392702004868	3,92167346818807	1,67671485878780
Erro máximo absoluto	2,72487526418241	2,01580538192229	1,39771007262909

Fonte: Autor, 2012.

Tabela C.5: Valores do pré-teste para o público-alvo da empresa de Tecnologia da Informação

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	3,02622660867734	3,10294204843090	3,73990210837565
Desvio Padrão	0,29667864873992	0,02861701301193	0,15236567795842
Erro quadrático	4,47077229594815	3,99951767966876	1,78748940701541
Erro máximo absoluto	2,37017683497195	1,92082255735586	1,43898255248413

Fonte: Autor, 2012.

Tabela C.6: Valores do teste final para o público-alvo da concessionária de veículos

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,51829289914485	3,02250428686351	4,17817622151892
Mediana	2,68970385043205	3,0096313818272	4,16837371946885
Desvio Padrão	0,512935928531575	0,16099548222671	0,21027488157152
Erro quadrático	6,58842935079273	4,01621474475985	0,73339340630657
Erro máximo absoluto	4,29289321881345	2,25090519486689	1,10999878996728

Fonte: Autor, 2012.

Tabela C.7: Valores do teste final para o público-alvo da farmácia

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,72963573369202	3,32068127990772	4,16949738536552
Mediana	2,64600085116561	3,3179504078975	4,16752361873625
Desvio Padrão	0,147193052602283	0,07014890837089	0,07464051857222
Erro quadrático	5,29076533872315	2,88258552613374	0,70938201610936
Erro máximo absoluto	2,63095050061707	1,80249532706978	0,92936793072424

Fonte: Autor, 2012.

Tabela C.8: Valores do teste final para o público-alvo do centro de estética

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,34453616447701	2,78409105459931	3,62957670599501
Mediana	2,57129193047477	2,76778163592573	3,5751997720566
Desvio Padrão	0,60337667250218	0,26825947461586	0,26055419423556
Erro quadrático	7,58753910731881	5,08242483378934	1,98427624808422
Erro máximo absoluto	3,88464492834958	2,61386590459572	1,67237958243060

Fonte: Autor, 2012.

Tabela C.9: Valores do teste final para o público-alvo da auto escola

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,73101748063608	3,10839734017566	3,77322186645950
Mediana	2,64027311202197	3,08238118840752	3,72156920882292
Desvio Padrão	0,522884325396304	0,06248922899534	0,12386706046236
Erro quadrático	5,53122759886416	3,65508921257155	1,55104160880340
Erro máximo absoluto	3,43198204536896	1,95630114015075	1,36864855201893

Fonte: Autor, 2012.

Tabela C.10: Valores do teste final para o público-alvo da empresa de Tecnologia da Informação

Medida\Cenário	Randômico	Georreferenciado	Abordagem Proposta
Média	2,72916958420082	3,15103065133812	3,79206134146748
Mediana	2,8050357265917	3,1518137870842	3,75831958207367
Desvio Padrão	0,457626040089929	0,07336122928439	0,12939465681505
Erro quadrático	5,50179423244803	3,49383865801428	1,50563665351567
Erro máximo absoluto	3,57627091614477	1,94499404856653	1,37725595535179

Fonte: Autor, 2012.

