Obsolescência de Informação

Sirley Ambrosia Vitorio

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA A
OBTENÇÃO DO GRAU DE MESTRE
EM
CIÊNCIA DA COMPUTAÇÃO

Área de Concentração: Ciência da Computação
Orientador: Prof. Dr. Flávio Soares Corrêa da Silva

— São Paulo, novembro de 2000 —

Obsolescência de Informação

Este exemplar corresponde à redação final da dissertação apresentada por Sirley Ambrosia Vitorio, devidamente corrigida e aprovada pela comissão julgadora.

Banca examinadora

Prof. Dr. Flávio Soares Corrêa da Silva(Presidente) IME-USP

Prof. Dr. Marcelo Finger IME-USP

Prof. Dr. Hugo Pietrantônio EP-USP

— São Paulo, novembro de 2000 —

OBSOLESCÊNCIA DE INFORMAÇÃO

por Sirley Ambrosia Vitorio

RESUMO

O tratamento de incerteza de informação é um tema muito estudado em Inteligência Artificial e conseqüentemente encontramos grande número de trabalhos na literatura.

Curiosamente, o tratamento de informações cuja incerteza varia com o tempo, não é facilmente encontrado na literatura. A queda gradual na confiabilidade da informação com o tempo é denominada *Obsolescência de Informação*, sendo encontrada freqüentemente em muitos problemas, como por exemplo, em finanças, economia, sociologia e demografia.

Nossa motivação para pesquisa surgiu a partir do projeto *SIDAM*¹ - *Sistemas de Informação Distribuídos para Agentes Móveis*, que fornece informações sobre o tráfego de automóveis em grandes cidades. Neste trabalho nós discutimos o fenômeno da Obsolescência de Informação e propomos um modelo para tratar com informações que perdem a confiabilidade com o tempo, apresentando uma proposta de implementação para este modelo.

¹financiado pela FAPESP, processo nr 98/06138-2

OBSOLESCEMENT INFORMATION

by Sirley Ambrosia Vitorio

ABSTRACT

Reasoning with uncertain information is a central issue in Artificial Intelligence and Information Systems development. Consequently, a vast amount of studies devoted to this issue can be found in the literature.

Curiously, reasoning with information whose uncertainty varies along time cannot be found so much in the literature. The gradual fall in the information reliability with the time is called *Obsolescent Information*.

Dynamic reliability is a feature found rather frequently in many problems, as in e.g. finance and economics, sociology, and demographics.

Our motivating case study is the information system under development within the project *SIDAM²-Distributed Information Systems for Mobile Agents, to provide information about urban automotive traffic for large cities*.

In the present research we discuss the issue of Obsolescent Information and propose a model for information that becomes progressively unreliable with time, introducing one propose of implemation of the model.

²supported by FAPESP grant no. 98/06138-2.

Ao meu filho Rafael.

AGRADECIMENTOS

À Deus, pela possibilidade de realização deste trabalho.

Ao Prof. Dr. Flávio Soares Corrêa da Silva, pela orientação e paciência demonstradas durante a realização deste trabalho.

A Roberto C. Araújo, pela preciosa ajuda na manipulação do $\text{\LaTeX}$ e interesse demonstrado durante a realização deste trabalho.

A Paulo Dantas, pelo incentivo e encorajamento demonstrados durante a elaboração deste trabalho.

A Tadeu Leite Duarte e à CET, pela gentileza de fornecer os dados utilizados nos testes do capítulo 5.

À minha família, por todo o incentivo, apoio e carinho.

Ao CNPq, pelo apoio financeiro concedido.

Enfim, agradeço a todos que, direta ou indiretamente, participaram do momento especial da realização deste trabalho.

Sumário

1	Introdução	13
2	Conceitos Preliminares	17
2.1	Introdução	17
2.2	Estatística	20
2.2.1	Probabilidade	23
2.2.2	Distribuição Normal	24
2.3	Funções	28
2.4	Métodos Numéricos	29
3	Um modelo para Obsolescência de Informação	31
4	Aplicação do Modelo	35
4.1	Proposta de Implementação do Modelo para a Aplicação . . .	39
4.2	Detalhamento das Funções	45
5	Testes e Resultados Obtidos	49
5.1	Testes	50
5.2	Resultados	64
6	Discussão e Trabalhos Futuros	67

Lista de Figuras

2.1	Esquema para solução de problemas	17
2.2	Refinamento do esquema para solução de problemas	18
2.3	Gráfico de densidade de probabilidade da função $f(x)$	25
2.4	Área sob a função de densidade de probabilidade no intervalo $[a, b]$	26
2.5	Gráfico de distribuição padrão	27
2.6	Gráfico de distribuição normal reduzida	28
3.1	Obsolescência de Informação: Fixado período de tempo	33
3.2	Obsolescência de Informação: Fixado valor de probabilidade	34
4.1	Fluxo de dados para o primeiro exemplo de consulta	41
4.2	Fluxo de dados para o segundo exemplo de consulta	41
4.3	Fluxo de informações entre o módulo modelo e as funções para dados externos e consultas.	42
4.4	Fluxo de informações entre as funções usadas na implementação do modelo.	44
5.1	Gráfico de distribuição normal de probabilidade	51
5.2	Curvas limitantes superior e inferior de probabilidade	51
5.3	Gráfico de distribuição normal de probabilidade para $t_1 > t_0$	52
5.4	Gráfico de OI	54
5.5	Gráfico de OI	55

5.6	Gráfico da variação de E_p	56
5.7	Gráfico de OI típico	59
5.8	Gráfico de OI típico	60
5.9	Gráfico de OI Atípico	61
5.10	Gráfico de OI Atípico	62
5.11	Gráfico de OI Oscilante	63

Lista de Tabelas

4.1	Tabela de discretização da velocidade	37
4.2	Tabela de discretização da velocidade e probabilidades estimadas	38
4.3	Tabela exemplo com valores de probabilidades estimada para cada classe de velocidade	39
4.4	Tabela das principais funções da proposta de implementação da aplicação	43
4.5	Tabela de tipos de dados da funções	47
5.1	Tabela de trechos da rua da Consolação	50
5.2	Tabela de classificação da variação da OI	56
5.3	Tabela dos valores de Ep utilizados para testes	64
5.4	Tabela de percentual de acertos para os diferentes tipos de gráficos	64

Capítulo 1

Introdução

O tratamento de incerteza de informação é um tema muito estudado em Inteligência Artificial e conseqüentemente encontramos grande número de trabalhos na literatura (veja ex. [2, 4, 6, 7, 8, 17]).

Curiosamente, o tratamento de informações cuja incerteza varia com o tempo, não é facilmente encontrado na literatura. A queda gradual na confiabilidade da informação com o tempo é denominada *Obsolescência de Informação (OI)* sendo encontrada freqüentemente em problemas de diferentes áreas, por exemplo:

finanças e economia: analistas da bolsa de valores precisam trabalhar com informações que devem ser continuamente atualizadas, sendo a sua confiabilidade relevante para os resultados de suas aplicações. Devido à característica competitiva deste negócio, as informações devem ser dinamicamente atualizadas para a avaliação de diferentes analistas.

sociologia e demografia: censo populacional é claramente dinâmico. A confiabilidade das informações coletadas a alguns dias atrás diminui com o tempo, pois a cada minuto ocorrem várias mortes e nascimentos, necessitando de constantes atualizações.

Nossa motivação para a pesquisa surgiu a partir do projeto *SIDAM – Sistema de Informações Distribuído para Agentes Móveis* (veja ex. [9, 10]). O objetivo do projeto SIDAM é fornecer uma infraestrutura que possibilite

a construção de sistemas de informações distribuídos complexos mantidos e consultados por agentes móveis. Um cenário típico para a aplicação desta infraestrutura é um sistema de informação para o tráfego de veículos urbanos em grandes cidades, no qual o motorista deseja tomar decisões dinâmicas sobre sua rota, para evitar congestionamentos.

As informações a serem tratadas em nossa pesquisa apresentam três características fundamentais:

Variação Temporal- em geral, os dados mudam com o tempo, necessitando atualizações.

Confiabilidade- existe um grau de incerteza associado a cada fonte de informação e ao modo de aquisição.

Validade- o valor da informação pode envelhecer, perdendo gradativamente a sua significância.

Estas características apresentam grande importância nos resultados obtidos com o tratamento destas informações. Desta forma, necessitam ser analisadas e modeladas para garantir a integridade de qualquer sistema que utilize informações cuja incerteza esteja associada.

Neste trabalho, nós analisamos e discutimos o fenômeno da OI.

Para o tratamento deste fenômeno, estaremos recebendo, a cada intervalo de tempo, estimativas estatísticas de valores de variáveis referentes às informações tratadas.

A perda de confiabilidade de informação, será analisada durante o período em que a informação não é atualizada. O intervalo de tempo é um valor não negativo, podendo representar segundos, minutos, dias, etc. Esta representação depende do tipo de informação a ser analisada.

As informações podem ser coletadas por diferentes fontes com grau de confiabilidade diferentes. Por exemplo:

Os dados referentes à velocidade do tráfego em uma grande cidade podem ser coletados por sensores de pavimento, radares, câmeras, helicópteros, pedestres, etc. Cada uma destas fontes possuem um grau de confiabilidade diferente que deve ser considerado.

A partir da análise das estimativas recebidas, propomos um modelo matemático para o tratamento da OI. Este modelo trabalha com o conceito de distância entre funções de probabilidade que nos permite estimar a queda da confiabilidade das informações tratadas.

Finalmente apresentamos uma proposta de implementação para este modelo.

Este trabalho está organizado da forma a seguir.

No capítulo 2 nós apresentamos resumidamente alguns conceitos preliminares utilizados em nosso modelo. Nós abordamos os conceitos básicos de estatística, probabilidade, funções, e apresentamos o modelo numérico adotado em nossa pesquisa.

No capítulo 3 nós apresentamos uma proposta de um modelo matemático para o tratamento da Obsolescência de Informação.

No capítulo 4 nós apresentamos uma aplicação para o modelo e uma proposta de implementação.

No capítulo 5 nós apresentamos alguns testes realizados com dados reais utilizando o nosso modelo proposto e a análise dos resultados obtidos.

Finalmente no capítulo 6 nós apresentamos uma discussão e proposta para trabalhos futuros.

Capítulo 2

Conceitos Preliminares

Neste capítulo apresentamos um resumo de alguns conceitos preliminares utilizados em nosso modelo.

2.1 Introdução

A maioria dos problemas matemáticos são originários da necessidade de resolver problemas da natureza. Isto porque fenômenos da natureza podem ser descritos através do uso de modelos matemáticos (veja ex. [5]).

Esquematicamente, podemos representar as etapas da solução de um problema conforme figura 2.1.

Nesta figura, observa-se numa primeira etapa, a obtenção de um modelo matemático que represente, da maneira mais conveniente, o problema específico que se quer estudar. A segunda etapa consiste em encontrar a solução do modelo proposto.

Figura 2.1: Esquema para solução de problemas

Figura 2.2: Refinamento do esquema para soluo de problemas

O modelo matemtico  construído utilizando teorias fsicas, qumicas, econmicas, etc. Para um mesmo problema, podemos ter modelos matemticos diferentes, dependendo do enfoque desejado. Em geral, ele contm simplificaes da realidade que nos levam a um problema matemtico solvel.

Podemos utilizar alguns mtodos numricos para obter a soluo numrica do modelo.

Podemos esquematizar a soluo de um problema mais detalhadamente conforme figura 2.2.

Nesta figura observa-se, em uma primeira etapa, que a construo do modelo matemtico  obtido a partir de um problema real e do levantamento de dados para este problema. A segunda etapa consiste na escolha do modelo numrico apropriado para se chegar  uma soluo. A prxima etapa

constitui a implementação computacional deste modelo, seguida da análise dos resultados obtidos. Finalmente, caso seja necessário, a reformulação do modelo ou escolha de um novo método numérico.

Seguindo este esquema o nosso trabalho esta organizado da seguinte forma:

Problema Real: consiste em como tratar informações que perdem a validade com o tempo.

Levantamento de dados: utilizaremos dados fornecidos pela *Companhia de Engenharia de Tráfego de São Paulo (CET/SP)*, no capítulo 5 especificaremos como estes dados foram obtidos e utilizados.

Modelo Matemático: utilizaremos conceitos de probabilidade, funções de probabilidade, distância e vizinhança entre funções, máximos e mínimos de funções. Estes conceitos estarão sendo abordados neste capítulo. No capítulo 3 estaremos introduzindo e explicando o modelo.

Modelo Numérico: para a solução do modelo estaremos escolhendo um modelo numérico apropriado que será abordado na seção 2.4.

Implementação Computacional: estaremos apresentando uma proposta de implementação, baseada no modelo matemático que será abordada no capítulo 4.

Análise dos Resultados: será abordada no capítulo 5.

Como foi comentado acima, alguns dos conceitos matemáticos que estaremos utilizando em nosso modelo são os de probabilidade e estatística. Desta forma, estaremos apresentando um resumo conceitual deste assunto na próxima seção.

2.2 Estatística

Estatística pode ser entendida como o estudo dos processos de obtenção, coleta, organização e análise de um conjunto de dados relevantes e referentes a qualquer fenômeno numericamente quantificável, sobre uma população, coleção ou conjunto de seres (veja ex. [15]).

Seu objetivo é permitir inferências (previsões, decisões) com base nos dados coletados e processados. Para se medir a incerteza de um processo, são utilizados cálculos de probabilidade. A seguir, apresentamos algumas terminologias:

População ou Universo: *é o conjunto dos seres, objetos ou informações cujas propriedades pretendemos estudar;*

Amostra: *é um subconjunto ou parte da população;*

Dados ou Variáveis: *informações ou dados característicos dos fenômenos ou populações, são classificados em qualitativos e quantitativos;*

Variáveis quantitativas: *podem ser expressas em termos numéricos, são divididas em contínuas e discretas:*

variáveis contínuas: *podem assumir um número infinitamente grande de valores, correspondentes aos pontos de um intervalo de reta de números reais, exemplos: medidas de tempo, comprimento, espessura, área, volume, peso, velocidade;*

variáveis discretas: *aquelas que presumem um número contável de valores, exemplos: número de alunos por sala, número de carros em um engarrafamento numa determinada área, número de parafusos defeituosas em uma amostra de 100 unidades.*

Variáveis qualitativas: *qualificadas segundo alguma de suas características típicas e não exclusivamente contáveis.*

Os dados coletados podem ser organizados na forma de listas, tabelas ou gráficos. As tabelas e os gráficos, apresentam os dados coletados. Esta forma é suficiente para análise quando os dados já vêm (ou podem ser) agrupados

num pequeno número de categorias. Quando isto não acontece ou quando o número de valores individuais é relativamente grande, estes devem ser agrupados em classes, categorias ou intervalos convenientemente escolhidos.

O método de tabulação ou agrupamento dos dados em classes, categorias ou intervalos é chamado de *Distribuição de Frequência*. Às tabelas dos dados agrupados desta forma denominam-se *Tabelas de Frequências*.

Existem alguns critérios sugeridos para se definir o número de intervalos (veja ex. [15]):

critério da raiz: por este critério, o número de intervalos (k) é igual à raiz quadrada do número de elementos observados (N): $k = \sqrt{N}$;

fórmula de Sturges: segundo Sturges o número de intervalos é dados pela expressão: $k = 1 + 3.3 \times \log N$;

fórmula de Milone/Angelini: é expressa pela fórmula: $k = 2 \times \ln N$.

Apesar de existirem estes critérios, de modo geral, pode-se dizer que a forma de agrupamento, e conseqüentemente o número de agrupamentos é arbitrário e depende, entre outras coisas, da natureza dos dados, preferências individuais e das necessidades e objetivos do usuário da informação (veja ex. [15]).

A frequência com que um determinado subgrupo de dados aparece na distribuição da tabela pode ser definida em termos absolutos ou em termos relativos. A frequência pode ser: absoluta, relativa ou acumulada.

Uma das primeiras medidas descritivas importantes é uma *medida de tendência central*, isto é, uma medida de ponto central da distribuição. Um das mais comuns e úteis é a *média aritmética* de um conjunto de medidas.

Média Aritmética: é a razão entre o somatório dos valores do conjunto considerado e o seu número de elementos. Dado um conjunto X de n elementos $x_1, x_2, \dots, x_n$ a média μ pode ser calculada pela fórmula:

$$\sum_{i=1}^n \frac{x_i}{n}.$$

Localizado o centro de uma dispersão de dados, o próximo passo é verificar a dispersão desses dados. A medida de dispersão mais simples é a amplitude.

Amplitude: *diferença entre o maior e o menor valor de certo conjunto de dados, pode ser referida ao total de dados ou a uma classe em particular.*

Variância: *dado um conjunto de n medidas $x_1, x_2, \dots, x_n$ a variância é definida como sendo a média dos quadrados dos desvios das medidas em relação a sua média μ . A notação usada é σ^2 e pode ser calculada pela fórmula:*

$$\sum_{i=1}^n \frac{(x_i - \mu)^2}{n}.$$

Desvio Padrão: *dado um conjunto de medidas $x_1, x_2, \dots, x_n$ é igual à raiz positiva da variância desse conjunto. A notação usada é σ .*

2.2.1 Probabilidade

Ainda não existe um consenso com relação à classificação da teoria da probabilidade. Alguns a classificam como um ramo da estatística, outros, como um ramo da matemática da qual a estatística se vale. Não estamos interessados nesta discussão, mas sim, na utilização dos recursos da estatística e probabilidade para o tratamento da incerteza. A *teoria da probabilidade* estuda os fenômenos ou situações em que os resultados possíveis dos experimentos são previamente conhecidos mas incertos quanto a sua ocorrência ou não. Nesse estudo, o que se busca é a quantificação da (in)certeza apresentado por certo resultado particular nas repetições futuras do mesmo experimento. A seguir apresentamos algumas terminologias importantes:

Experimento: *é o fato ou fenômeno que está sendo estudado;*

Experiência: *é o processo pelo qual uma observação ou medida é obtida;*

Espaço Amostral: *é o conjunto de todos os resultados possíveis de certo experimento;*

Evento elementar: *é cada um dos resultados possíveis do espaço amostral do experimento.*

Dado um evento A , a probabilidade de ocorrer este evento é dado pela fórmula $P(A) = \frac{n}{N}$, sendo n o número de vezes que o evento A ocorreu e N o número de vezes que a experiência foi repetida.

A verdade e a falsidade sobre um fenômeno em estudo podem ser afirmadas, quando este está sujeito a alguma lei que permite prever seus resultados a partir das informações disponíveis. Os modelos que explicam estes fenômenos são chamados de *modelos determinísticos*, resultantes de anos ou até séculos de estudos. Por outro lado, se o fenômeno em estudo não está sujeito (ou não conhecemos) a uma lei que permita prever seus resultados, o modelo possível para seu estudo é chamado *modelo probabilístico*. O modelo probabilístico é aquele que é aplicável aos fenômenos cujos resultados exatos em cada experimento possível são desconhecidos e imprevisíveis. O estudo de tais modelos é o objetivo da teoria de probabilidades. Sua aplicação só é possível nas situações onde há incerteza quanto à ocorrência ou não dos eventos estudados.

2.2.2 Distribuição Normal

Conforme foi visto na seção anterior, uma *variável aleatória discreta* é aquela que presume um número contável de valores. Exemplos: número de falhas em um avião durante um determinado período de vôo, número de parafusos defeituosos em uma amostra com N unidades, sendo N inteiro. A distribuição de probabilidade de uma variável aleatória discreta constitui uma fórmula, tabela ou gráfico que possibilita calcular a probabilidade associada a cada valor dessa variável. Uma *variável aleatória contínua* é aquela que pode assumir um número infinitamente grande de valores, correspondentes aos pontos de um intervalo de uma reta de números reais. Exemplos típicos de variáveis aleatórias contínuas:

1. altura de um homem
2. tempo de vida de uma célula humana
3. medidas de laboratório

O modelo probabilístico para a distribuição de frequências de uma variável aleatória contínua envolve a escolha de uma curva denominada *Distribuição de Probabilidade de Variável Aleatória* ou *Função da Densidade de Probabilidade*.

Embora essas distribuições possam pressupor uma variedade de formas, é interessante notar que uma grande quantidade de variáveis aleatórias observadas na natureza apresentam uma distribuição de frequência aproximadamente bem-comportada, ou, estatisticamente, aproximadamente normal.

Uma variável aleatória contínua X tem distribuição normal de probabilidade se a sua função de densidade de probabilidade é dada por (veja ex. [14]):

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}, \text{ para } -\infty < x < \infty$$

Onde e e π representam números irracionais, cujos valores aproximados são, respectivamente: 2,7183 e 3,1416. μ é o valor da média da população considerada e σ o desvio padrão.

Figura 2.3: Gráfico de densidade de probabilidade da função $f(x)$

O gráfico de $f(x)$ pode ser representado como na figura 2.3.

Nesta figura observa-se a curva de distribuição normal da função $f(x)$, e os pontos referentes à média(μ) e desvio padrão(σ).

A equação para a função de densidade é obtida de modo que a área sob a curva represente probabilidade, sendo assim, o valor total dessa área é igual a 1.

As principais características dessa função são (veja ex. [15]):

1. o ponto máximo de $f(x)$ é o ponto $X = \mu$
2. Os pontos de inflexão da função são: $X = \mu + \sigma$ e $X = \mu - \sigma$
3. A curva é simétrica com relação à μ
4. $E(X) = \mu$ e $Var(X) = \sigma^2$

Além disso, devemos notar que, a equação de distribuição normal é dependente dos valores numéricos de μ e σ . Assim, com diferentes valores desses parâmetros, pode-se obter uma quantidade infinitamente grande de distribuições normais. Na realidade temos uma família de distribuições normais.

Uma mudança no valor de μ não muda a forma da figura, o que ocorre é um deslocamento da figura para a direita ou para a esquerda, dependendo do novo valor ser maior ou menor que o anterior.

Figura 2.4: Área sob a função de densidade de probabilidade no intervalo $[a, b]$

Uma mudança no valor de σ , no entanto, muda a forma da figura, embora ela seja sempre simétrica em relação à média, para valores pequenos de σ ela tem uma forma mais alongada e mais estreita, enquanto que para valores grandes ela será mais longa e achatada.

A probabilidade de uma variável aleatória contínua assumir um valor no intervalo $[a, b]$ é igual à área sob a curva de sua função de densidade de probabilidade entre os pontos a e b . Ver figura 2.4.

Nesta figura observa-se a área sob a curva de distribuição normal, entre os pontos do intervalo $[a, b]$ que é igual ao valor de probabilidade de uma variável aleatória assumir um valor neste intervalo.

Para calcular as áreas sob a curva normal, devemos fazer (ver ex. [15]):

$$P(a \leq x \leq b) = \int_a^b \frac{1}{\sigma \sqrt{2\pi}} e^{-\left(\frac{x-\mu}{2\sigma}\right)^2} dx$$

que apresenta um grau relativo de dificuldade.

Uma tabela que indique os valores de cada uma das possíveis distribuições é impraticável, no entanto, existe uma tabela que pode ser aplicada a qualquer curva normal. Uma vez que a curva normal é simétrica em relação à média, podemos simplificar uma tabela de áreas, listando apenas as áreas compreendidas entre a média e um número específico z , de desvio padrão, à direita de μ . As áreas à esquerda da média, poderão ser obtidas, usando-se a simetria.

Figura 2.5: Gráfico de distribuição padrão

Inicialmente nós convertemos a média X na chamada *Média Padrão* z :

$$z = \frac{x - \mu}{\sigma}$$

Teremos os pontos $z_1 = \frac{a - \mu}{\sigma}$ e $z_2 = \frac{b - \mu}{\sigma}$, conforme a figura 2.5.

Nesta figura observa-se o gráfico de distribuição normal reduzida com a conversão da média para a média padrão z .

Os pontos $\mu - \sigma$ e $\mu + \sigma$ transformam-se em -1 e $+1$. Ver figura 2.6.

Nesta figura observa-se no gráfico de distribuição normal reduzida os pontos $\mu - \sigma$ e $\mu + \sigma$ transformados em -1 e $+1$

Z também tem distribuição normal, é chamada *variável normal reduzida ou padronizada*. Através desta transformação podemos calcular a probabilidade de qualquer variável contínua usando a *Tabela de Distribuição Normal* (veja ex. [14]).

Figura 2.6: Gráfico de distribuição normal reduzida

2.3 Funções

Noções elementares sobre funções são amplamente encontradas na literatura(ver ex [1, 11, 12, 13]). A seguir apresentamos resumidamente alguns conceitos envolvendo funções que estaremos utilizando: distância, vizinhança, máximos e mínimos. No capítulo 3 apresentamos a sua utilização e extensão.

Distância(ver ex. [1, 13, 18])

Dado dois pontos $A = (x_1, y_1)$ e $B = (x_2, y_2)$ temos que a distância entre A e B , denotada $d(A, B)$ é calculada segundo a fórmula:

$$d(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

Vizinhança(ver ex. [13])

Dado um ponto $P_0 = (x_0, y_0)$ e um número $\delta > 0$, chama-se vizinhança δ de P_0 , que se indica $V_\delta(P_0)$ ao conjunto de pontos $P = (x, y)$ cuja distância a P_0 é menor que δ .

Máximos e Mínimos(ver ex. [1, 13])

Seja f uma função definida no intervalo $[a, b]$. Dizemos que f tem um máximo local; no ponto x_0 do intervalo, se, para algum $q > 0$, $f(x) \leq f(x_0)$ para todo x no intervalo $[a, b]$ tal que $q < x < x_0 + q$.

O mínimo local é definido da mesma forma trocando o sinal $\leq$ por $\geq$. Dizemos que f tem um máximo absoluto M se $f(x_0) = M$, para pelo menos um x_0 no intervalo $[a, b]$ e se $f(x) \leq M$ para todo x neste intervalo. O mínimo absoluto é definido de maneira análoga.

2.4 Métodos Numéricos

Por método numérico entende-se como um conjunto de regras escritas sob a forma de uma sequência de operações elementares que nos levam à uma solução do problema.

A solução obtida pelo método numérico pode diferir da solução do problema real. Podemos enumerar algumas fontes de erro que levam a essa diferença.

1. Simplificações no modelo matemático.
2. Erro de truncamento
3. Erro de arredondamento cometido durante os cálculos
4. Erro nos dados

A idéia central dos métodos numéricos é partir de uma aproximação inicial para a raiz de uma função e em seguida “refiná-la” através de um processo iterativo.

Por isto os métodos contam de duas fases:

1. Localização ou isolamento da raiz, que consiste em obter um intervalo que contém a raiz.
2. Refinamento, que consiste em, escolhidas aproximações iniciais no intervalo encontrado na fase I, melhorá-las sucessivamente até se obter uma aproximação para a raiz dentro de uma precisão pré-fixada.

Um dos processos iterativos para calcular um valor aproximado de uma raiz, supondo já encontrado um intervalo contendo somente esta raiz é o *Método da Dicotomia ou Bissecção*.

Supondo que uma raiz da função f esteja isolada no interior do intervalo $[a, b]$ e portanto $f(a) * f(b) < 0$, o processo consiste em dividir o intervalo dado ao meio, e por aplicações do teorema de Bolzano (veja ex.[3]) aos sub-intervalos:

$$\left[a, \frac{a+b}{2}\right]$$

e

$$\left[\frac{a+b}{2}, b\right],$$

determinar qual deles contém a raiz.

O processo é repetido para o novo sub-intervalo até que se obtenha uma precisão pré-fixada.

Em nosso modelo nós utilizaremos a idéia básica do método de Dicotomia. Inicialmente consideramos o intervalo de tempo $[a, b]$. Para cada novo valor de $t = \frac{a+b}{2}$, utilizamos duas funções de nossa proposta de implementação para verificar se este valor de tempo satisfaz ao resultado esperado. Enquanto não obtemos o resultado com a precisão pré-fixada o processo é repetido. Sua utilização é detalhada no capítulo 4.

Capítulo 3

Um modelo para Obsolescência de Informação

Nós consideramos que a informação está sendo “discretizada”, isto é, estará sendo classificada para constituir um domínio finito de “valores relevantes”.

Cada um destes valores será chamado de *evento atômico*. O que estaremos recebendo a cada intervalo de tempo são *probabilidades atômicas estimadas* para um conjunto finito de eventos atômicos.

Um conjunto de eventos atômicos é denotado como $E = \{e_1, \dots, e_n\}$. Uma probabilidade atômica estimada é uma função real $\mu_e : E \rightarrow [0, 1]$, tal que $\sum_E \mu_e(e_i) = 1$.

Cada *evento* é algum subconjunto de E . Uma *probabilidade estimada* para um conjunto de eventos baseada na probabilidade atômica estimada μ_e é uma função real $\mu : 2^E \rightarrow [0, 1]$ tal que $\mu(\mathcal{E}) = \sum_{e_i \in \mathcal{E}} \mu_e(e_i)$, $\mathcal{E} \in 2^E$.

Dados duas probabilidades atômicas estimadas $\mu_{e1}, \mu_{e2} : E \rightarrow [0, 1]$, a *distância* entre as correspondentes probabilidades estimadas é definida como $D(\mu_1, \mu_2) = \max\{|\mu_{e1}(e_i) - \mu_{e2}(e_i)| : e_i \in E\}$. A δ -vizinhança da probabilidade estimada $\mu : 2^E \rightarrow [0, 1]$ ($\delta(\mu)$) é definida como:

$$\delta(\mu) = \{(\mu' : 2^E \rightarrow [0, 1]) : D(\mu, \mu') \leq \delta\}$$

Claramente, para $\delta \geq 1$, o δ -vizinhança para alguma probabilidade esti-

mada é o conjunto de todas as possíveis probabilidades estimadas para um dado conjunto de eventos 2^E .

Estaremos trabalhando com valores de tempo, reais e lineares, no eixo $T = [0, \infty)$, sendo $x \in [a, b)$ se $a \leq x < b$. O tempo pode ser referido como horas, dias, meses, etc conforme a natureza do problema.

Estaremos analisando o tratamento da perda de confiabilidade da informação durante o intervalo de tempo em que as informações não são atualizadas. Não havendo atualização, o processo de obsolescência prossegue, sendo que a informação torna-se cada vez menos confiável.

Nós definimos uma *função de vizinhança dependente do tempo* $\delta : T \rightarrow [0, 1]$ como alguma função não decrescente tal que $\delta(0) = 0$. Isto porque no momento do recebimento das informações ainda não ocorreu OI e portanto a perda de confiabilidade é zero.

Dado uma probabilidade estimada μ_0 para um instante inicial $t_0 = 0, t \in T$, nós assumimos que a probabilidade estimada para $t > 0$ pertencerá a δ -vizinhança de μ_0 , onde $Ep = \delta(t)$. Ep estará quantificando a perda de confiabilidade após um intervalo de tempo $t > 0$ ou seja a OI.

Como uma consequência, dado um evento $\mathcal{E}$, nós não podemos acessar a probabilidade estimada $\mu(\mathcal{E})$ quando $t > 0$, por que nós não sabemos qual a probabilidade estimada é a atual – tudo o que sabemos é que $\mu \in \delta(t)(\mu_0)$.

O que nós podemos dizer sobre a probabilidade estimada para $\mathcal{E}$ é que, para cada evento atômico $e \in \mathcal{E}$, correspondem probabilidade atômica estimada que pertencem ao intervalo

$$[\max\{0, \mu_0(e) - \delta(t)\}, \min\{1, \mu_0(e) + \delta(t)\}].$$

Vários tipos de consultas podem ser aplicados à este modelo, exemplos:

Consulta1: Dado um evento de interesse $\mathcal{E}$ e um intervalo de tempo $t \geq 0$, nós podemos estar interessados em determinar o intervalo ao qual a probabilidade estimada $\mu(\mathcal{E})$ pertence.

Consulta2: Dado um evento de interesse $\mathcal{E}$ e um valor de probabilidade, nós podemos estar interessados em determinar o tempo ao qual este valor de probabilidade é válido.

Desta consulta podemos obter duas consultas derivadas:

Figura 3.1: Obsolescência de Informação: Fixado período de tempo

Consulta3: Dado um evento de interesse $\mathcal{E}$ e um valor de probabilidade, nós podemos estar interessados em determinar o tempo no qual este valor de probabilidade é o mínimo.

Consulta4: Dado um evento de interesse $\mathcal{E}$ e um valor de probabilidade, nós podemos estar interessados em determinar o tempo no qual este valor de probabilidade é o máximo.

Consulta5: Dado um intervalo de tempo t , nós podemos estar interessados em determinar qual(is) evento(s) com maior probabilidade de ocorrer neste instante.

Consulta6: Dado um intervalo de tempo t e um valor de probabilidade, nós podemos estar interessados em determinar qual(is) evento(s) possuem esta probabilidade.

Nas figuras 3.1 e 3.2 observa-se os dois primeiros exemplos de consultas. Na figura 3.1 observa-se o dinamismo dos intervalos para probabilidades estimadas, mostrando a evolução esperada para um intervalo após um período de tempo fixado. Na figura 3.2 observa-se o dinamismo de intervalos de probabilidade estimadas, mostrando o período de tempo necessário para o valor de probabilidade fixada ser válida.

Figura 3.2: Obsolescência de Informação: Fixado valor de probabilidade

Capítulo 4

Aplicação do Modelo

Neste capítulo estaremos apresentando uma aplicação do nosso modelo utilizando dados específicos. Para esta aplicação estaremos utilizando dados coletados na cidade de São Paulo, fornecidos pela CET. As informações são obtidas por sensores, sobre as quais são estimados velocidade média esperada e desvio padrão do tráfego em intervalos fixos de quinze minutos.

As informações são coletadas a cada dois minutos por cada sensor, normalmente localizados em cruzamentos estratégicos, mas não necessariamente sincronizados. Porém estas informações “pontuais”, obtidas a cada dois minutos, não são informativas. A CET/SP tem adotado um procedimento de acúmulo de informações, a cada quinze minutos, para produzir estatísticas de valores de variáveis neste período.

Atualmente a CET/SP utiliza o sistema *Infovel* (ver [20])—um método de avaliação de velocidade em tempo real baseado em dados obtidos por sistemas de controle semafóricos de tempo real, tais como SCOOT e ITACA (ver [16, 19]). Este sistema disponibiliza a velocidade média das vias de tráfego, calculada a partir de informações obtidas dos laços detetores dos semáforos em tempo real e dos radares fotográficos de fiscalização de velocidade. Disponibiliza a informação sobre os níveis de velocidade a cada dois minutos (instantânea) e em intervalos de quinze minutos (histórica) para um banco de dados obtido para velocidades médias.

O estudo da OI sobre o evento velocidade é muito interessante e apropriado, visto que a situação do tráfego nas grandes cidades têm influência

relevante em nosso dia a dia.

A velocidade apresenta as 3 características fundamentais que são tratadas em nossa pesquisa:

Variação Temporal: ao recebermos um valor de velocidade num instante de tempo t_0 , necessitamos da sua atualização pois com o passar do tempo estes valores já podem ter se modificado;

Confiabilidade: os valores referentes à velocidade do tráfego podem vir de fontes diferentes(sensores, câmeras, pedestres, etc) cada uma delas possui um grau de confiabilidade diferente que deve ser considerado;

Validade: com o passar do tempo, o valor recebido perde gradativamente a sua significância, ficando cada vez menos confiável.

É importante observar que para cada aplicação podem existir várias alternativas de informações a serem analisadas. Por exemplo, para o nosso caso específico, ao invés da velocidade média, poderíamos utilizar o desvio padrão de velocidades específicas. Utilizamos a velocidade média por ser esta, a priori, a informação disponível referente ao tráfego. Além disto consideramos que o sistema que utilizará o modelo não possui os dados históricos referentes à informação, mas sim os dados instantâneos que são substituídos a cada atualização. Um melhoramento desta análise será proposto no capítulo 6 como trabalhos futuros.

Seguindo o nosso modelo, o primeiro passo é a “discretização” da informação. Os valores referentes à velocidade são contínuos, sendo assim precisamos dividi-los em intervalos ou classes. Estes valores são dados em Km/h e podem variar de 0 a 200. No capítulo 2 vimos alguns critérios que podem ser usados para obtenção do número de intervalos. Temos os seguintes resultados:

critério da raiz: $k = \sqrt{200} = 14.142$;

fórmula de Sturges: $k = 1 + 3.3 \times \log 200 = 1 + 3.3 \times 2.301 = 8.6$;

fórmula de Milone/Angelini: $k = 2 \times \ln 200 = 2 \times 5.29 = 10.6$.

classes	intervalos
1	[0, 20)
2	[20, 40)
3	[40, 60)
$\vdots$	$\vdots$
10	[180, 200)

Tabela 4.1: Tabela de discretização da velocidade

O número de classes que podemos dividir varia entre 8 e 14, com tamanhos arbitrários. A fórmula de Milone/Angelini nos fornece um valor médio, o qual estaremos adotando. Estaremos discretizando a velocidade em 10 classes de comprimento fixos conforme a tabela 4.1.

Nesta tabela os valores da informação velocidade, são discretizados em 10 classes de tamanhos fixos.

A escolha de tamanhos fixos deve-se ao fato de desejarmos testar o modelo para informações genéricas, cujo comportamento não conhecemos a priori. Apesar de intuitivamente conhecermos algumas características do trânsito em São Paulo, que nos levariam a pensar em classes de tamanhos diferentes, o objetivo da aplicação é testar o modelo genérico e não a análise específica de um determinado dado. Esta análise é proposta no capítulo 6 como trabalhos futuros.

Cada uma dessas classes constitui um evento atômico. Neste caso o conjunto de eventos atômicos E é constituído pelas 10 classes de velocidade. μ_e é a probabilidade atômica estimada, tal que $\sum_E \mu_e(e_i) = 1$. Ver tabela 4.2.

Nesta tabela observa-se para cada classe (evento) um valor de probabilidade atômica estimada μ_e .

Conforme o modelo proposto para o tratamento da obsolescência de informação, a cada evento atômico está associado um valor de medida de probabilidade. Estamos trabalhando com variáveis aleatórias contínuas e para descrever o seu comportamento, assumimos que as informações probabilísticas fornecidas pelos sensores (velocidade do tráfego) obedecem ao Modelo de Distribuição Normal (veja ex. [14]).

No modelo probabilístico assumido, a probabilidade de uma variável ale-

classes	intervalos	μ_e
1	$[1, 20)$	μ_{e1}
2	$[20, 40)$	μ_{e2}
3	$[40, 60)$	μ_{e3}
4	$[60, 80)$	μ_{e4}
$\vdots$	$\vdots$	$\vdots$
10	$[180, 200)$	μ_{e10}

Tabela 4.2: Tabela de discretização da velocidade e probabilidades estimadas

atória contínua assumir valores no intervalo $[a, b)$ é igual à área sob a curva da função densidade de probabilidade entre os pontos a e b (ver ex. [14]). Assim, dados velocidade esperada e desvio padrão como acima, nós podemos estimar a probabilidade de cada classe que corresponde a um intervalo de velocidade. Assim, quando nos referimos a um evento, por exemplo velocidade igual a $70 Km/h$, estaremos nos referindo à classe à qual pertence esta velocidade.

Dado uma função de vizinhança dependente do tempo $\delta(t)$, definida como uma função não decrescente de acordo com o problema considerado, nós podemos estimar o intervalo de probabilidade para cada classe de velocidade em algum periodo de tempo t após uma dada probabilidade estimada gerada pelo processamento de informações fornecidas pelos sensores. Como exemplo temos a tabela 4.3.

Nesta tabela observa-se a discretização da velocidade e os valores de probabilidade estimada para cada classe no instante do recebimento destas informações ($t = 0$). Se quisermos analisar a obsolescência sobre o evento velocidade $= 70 Km/h$ após 5 minutos ($t_e = 5$), considerando por exemplo que $\delta(t) = 0.01 \times t$ teríamos $\delta(5) = 0.05$ resultando: $0.45 \leq \mu(70) \leq 0.55$

Com o passar do tempo, o intervalo de probabilidade aumenta, representando o aumento da obsolescência da informação, conseqüente da perda da confiabilidade com o tempo. O processo continua até que as informações são atualizadas e recomeçamos o processo, caso isto não ocorra chegaremos a um instante de tempo a partir do qual não podemos garantir a confiabilidade da informação ($0.0 \leq \mu_e(v) \leq 1.0$).

4.1. PROPOSTA DE IMPLEMENTAÇÃO DO MODELO PARA A APLICAÇÃO 39

classes	intervalos	μ_e
1	[0, 20)	0.01
2	[20, 40)	0.09
3	[40, 60)	0.15
4	[60, 80)	0.50
$\vdots$	$\vdots$	$\vdots$
10	[180, 200)	0.01

Tabela 4.3: Tabela exemplo com valores de probabilidades estimada para cada classe de velocidade

4.1 Proposta de Implementação do Modelo para a Aplicação

O nosso trabalho consiste em um modelamento para o tratamento de informações que perdem a confiabilidade com o tempo. O objetivo da implementação é o de mostrar a validade do modelo na prática e permitir que ele seja utilizado para consultas a respeito da confiabilidade (incerteza) da informação desejada.

A partir deste resultado as informações devem ser analisadas e incorporadas ao processamento de um sistema global para uma resposta final ao usuário (por exemplo: motorista). Sendo assim, estaremos recebendo como dados externos, informações da CET/SP e parâmetros das consultas feitas pelo sistema, retornando o resultado adequado ao tipo de consulta realizada. Considerando os dois primeiros exemplos de consultas, apresentados na seção 3, temos:

C1: Dado um evento de interesse $\mathcal{E}$ e um tempo $t \geq 0$, nós podemos estar interessados em determinar o intervalo no qual a probabilidade estimada $\mu(\mathcal{E})$ pertence.

Nós recebemos como dados externos:

1. velocidade esperada vel ;
2. desvio padrão dp ;
3. horário inicial H_0 ;

4. evento atômico v ; e
5. horário esperado He .

Os valores 1, 2 e 3 são fornecidos pela CET/SP e os valores 4 e 5 são fornecidos pela consulta feita ao módulo pesquisa. Os dados referentes ao horário estão no formato *hora:minuto*, o valor de t pode ser obtido pela diferença entre H_e e H_0 .

De forma similar, para uma consulta diferente, apenas os valores 4 e 5, referentes às consultas são modificados.

C2: Dado um evento de interesse $\mathcal{E}$ e um valor de probabilidade, nós podemos estar interessados em determinar o tempo no qual este valor de probabilidade é válido.

1. velocidade esperada vel ;
2. desvio padrão dp ;
3. Horário inicial H_0 ;
4. evento atômico v ; e
5. valor de probabilidade PE .

O fluxo de dados de C1 e C2 estão representados nas figuras 4.1 e 4.2 respectivamente.

Na figura 4.1 observa-se o fluxo de dados para o exemplo do primeiro tipo de consulta. A CET nos fornece velocidade média (vm), desvio padrão (dp) e o horário (H_0) a partir do qual estas informações são válidas. A partir destes dados acrescido da velocidade esperada (v) e horário esperado (He) fornecidos pela consulta, o módulo pesquisa pode retornar o intervalo de probabilidade $[Pi, Pf]$ válido.

Na figura 4.2 observa-se o fluxo de dados para o exemplo do segundo tipo de consulta. A CET nos fornece velocidade média (vm), desvio padrão (dp) e o horário (H_0) a partir do qual estas informações são válidas. A partir destes dados mais a velocidade esperada (v) e Probabilidade esperado (He), fornecidos pela consulta, o módulo pesquisa pode retornar o horário (He) a partir do qual estes valores são válidos

4.1. PROPOSTA DE IMPLEMENTAÇÃO DO MODELO PARA A APLICAÇÃO041

Figura 4.1: Fluxo de dados para o primeiro exemplo de consulta

Figura 4.2: Fluxo de dados para o segundo exemplo de consulta

Figura 4.3: Fluxo de informações entre o módulo modelo e as funções para dados externos e consultas.

Como vimos, vários tipos de consultas podem ser feitas ao módulo pesquisa, que serão direcionadas para funções específicas que irão manipular o nosso modelo de forma a obter resultados pertinentes à consulta realizada. Assim, estaremos trabalhando com funções que recebem dados externos e retornam valores resultantes da manipulação das funções que representam o nosso modelo e estão contidas no que chamamos módulo modelo. Ver figura 4.3.

Nesta figura observa-se a função *In_Cet* que recebe os dados de velocidade média, desvio padrão e horário inicial e os fornece ao módulo modelo. Temos as funções *Calc_OIn*, que irão manipular as funções internas do módulo modelo de forma a responderem à consulta efetuada.

Os dados fornecidos pela CET estarão armazenados em uma base de dados, disponíveis para consulta. Cada vez que necessitarmos destes dados, uma função chamada *In_Cet* é acionada e carrega estes dados deixando-os disponíveis para o sistema. Duas das funções de consulta são as referentes aos exemplos de consultas utilizados anteriormente. A função chamada *Calc_OI1* recebe o evento velocidade (v) e um horário (He) - composto por hora:minuto - e retorna o intervalo de probabilidade ($[Pi, Pf]$) desta velocidade no horário pedido. A função chamada *Calc_OI2* recebe o evento velocidade (v) e um valor de probabilidade (P) e retorna o horário no qual esta probabilidade é válida.

Na tabela 4.4, estão listadas as principais funções utilizadas em nosso

4.1. PROPOSTA DE IMPLEMENTAÇÃO DO MODELO PARA A APLICAÇÃO 43

In_Cet(vel,dp,H0);
Calc_Class(v, sz);
Calc_Prob(cl, sz, vel,dp);
Calc_Func(t);
Calc_InterProb(P0, Epi);
Calc_Time(P0,Pe);
Calc_OI1(v, He);
Calc_OI2(v, PE);

Tabela 4.4: Tabela das principais funções da proposta de implementação da aplicação

sistema e a sua utilização.

In_Cet: lê os dados de entrada fornecidos pela *CET – velocidade média esperada (vm), desvio padrão (dp), horário inicial* - e deixa-os disponíveis para serem utilizados pelo sistema.

Calc_Class: dados *velocidade v* e *tamanho da classe sz* (no nosso caso tamanho da classe é 20, retorna a que classe esta velocidade (*v*) pertence.

Calc_Prob: dados a *classe cl*, *tamanho da classe sz*, *velocidade média esperada (avs)* e *desvio padrão (sd)* retorna a *Probabilidade inicial estimada P0* para tempo=0 (horário inicial H_0).

Calc_Func: dado um período de tempo t , retorna o valor da função de vizinhança $\delta(t)$.

Calc_InterProb: dado a *Probabilidade inicial estimada P0* retornada por *Calc_Prob* e o valor da vizinhança *Epi* retornado por *Calc_Func*, retorna o intervalo de probabilidade $[Pi, Ps]$.

Calc_Time: dado a *Probabilidade inicial estimada P0* retornada por *Calc_Prob*, e um valor de probabilidade esperada *PE* , retorna o intervalo de tempo (te) tal que $PE \in [Pi, Ps]_i$.

Calc_OI1: dado *velocidade v* e um horário *He* , retorna o intervalo de probabilidade $[Pi, Ps]$ para a velocidade *v* no horário *He*.

Figura 4.4: Fluxo de informações entre as funções usadas na implementação do modelo.

Calc_OI2: dado a *velocidade* e um *valor de probabilidade PE* retorna o horário no qual esta probabilidade *PE* é válida.

Na figura 4.4 é apresentado o fluxo de dados entre estas funções.

4.2 Detalhamento das Funções

In_Cet: No momento em que esta função é ativada, ela consulta a base de dados da *CET/SP* e armazena os valores – *velocidade média esperada* (*vm*), *desvio padrão* (*dp*), *horário inicial* (*horário a partir do qual as informações anteriores devem ser consideradas válidas*)– em variáveis que estarão disponíveis para serem utilizados pelo sistema.

Calc_Class: Calcula o número da classe ao qual a *velocidade v* pertence que corresponde à parte inteira do resultado de v/sz , sendo *sz* o *tamanho da classe* considerada no problema.

Calc_Prob: A partir da *classe cl* e do *tamanho da classe sz* calculamos os valores dos limitantes da classe considerada. Obtidos estes valores utilizamos a *velocidade média esperada* (*vm*) e *desvio padrão* (*dp*) para o cálculo dos valores das variáveis normalizadas, que nos permitem encontrar o valor da *Probabilidade inicial estimada P0* consultando uma Tabela de Distribuição Normal, inclusa na implementação.

Dado uma classe *cl* de tamanho *sz* os limitantes desta classe são [*vi*,*vf*) obtidos pelas expressões:

$$vi = cl * sz$$

e

$$vf = (cl + 1) * sz - 1$$

Os valores das variáveis normalizadas correspondentes a estes limitantes são obtidos pelas expressões:

$$zi = (vi - vm)/dp$$

e

$$zf = (vf - vm)/dp$$

Esta função utiliza uma função interna *prob* que analisa os valores de *zi* e *zf* e encontra o valor de *P0* na Tabela de Distribuição Normal.

Calc_Func: recebe um período de tempo t calculado fora desta função e que corresponde à diferença, no nosso caso em minutos, entre o horário esperado He (horário para o qual estimamos o intervalo de probabilidade) e horário inicial $H0$ (armazenado pela *CET/SP*). A partir deste valor calcula o resultado da função de vizinhança $\delta(t)$.

Calc_InterProb: a partir da *Probabilidade inicial estimada* $P0$ retornada por *Calc_Prob* e o valor da vizinhança Epi retornado por *Calc_Func* calcula o intervalo de probabilidade $[Pi, Ps]$:

$$Pi = P0 - Epi$$

e

$$Pf = P0 + Epi$$

Calc_Time: dado a *Probabilidade inicial estimada* $P0$ retornada por *Calc_Prob*, e um valor de probabilidade esperada PE , retorna o intervalo de tempo (te) tal que $PE \in [Pi, Ps]_i$.

O valor de te é obtido através de processos iterativos de cálculo numérico. Utilizamos o método de dicotomia ou bissecção para gerar os intervalos de tempo juntamente com as funções *Calc_Func* e *Calc_InterProb* para auxiliarem no processo de verificação dos intervalos de probabilidade obtidos a partir dos períodos de tempo gerados.

Calc_OI1: dado *velocidade* v e um horário He , retorna o intervalo de probabilidade $[Pi, Ps]$ para a velocidade v no horário He . Para isto esta função manipula as funções do nosso modelo:

1. Utiliza a função *In_Cet*
2. Utiliza a função *Calc_Class*
3. Utiliza a função *Calc_Prob*
4. Utiliza a função *Calc_Func*
5. Utiliza a função *Calc_InterProb*

Calc_OI2: dado a *velocidade* e um *valor de probabilidade* PE retorna o horário no qual esta probabilidade PE é válida. Para isto esta função manipula as funções do nosso modelo:

<code>void In_Cet(float *avs, float *sd , hour *H0);</code>
<code>int Calc_Class(float v, int sz);</code>
<code>double Calc_Prob(int cl, int sz, float avs, float sd);</code>
<code>double Calc_Func(float t);</code>
<code>interp Calc_InterProb(double P0, double Epi);</code>
<code>float Calc_Time(double P0,double Pe);</code>
<code>interp Calc_OI1(float v, hour He);</code>
<code>hour Calc_OI2(float v, double PE);</code>

Tabela 4.5: Tabela de tipos de dados da funções
interp é uma estrutura composta por dois valores $\text{double}[P_{\text{inf}}, P_{\text{sup}}]$, e *hour* é uma estrutura para horário em horas:minutos

1. Utiliza a função *In_Cet*
2. Utiliza a função *Calc_Class*
3. Utiliza a função *Calc_Prob*
4. Utiliza a função *Calc_Time*

Tipos de Dados da Implementação

O programa foi implementado utilizando a linguagem C padrão. Sendo composto por um conjunto fixo de funções: *Calc_Class*, *Calc_Prob*, *Calc_InterProb*, *Calc_Func*, *Calc_Temp*, *In_Cet*, que representam o nosso modelo matemático.

As funções *Calc_OIn* (*Calc_OI1*, *Calc_OI2*,) aplicam as funções do nosso modelo de diferentes formas para chegar ao resultado da consulta. Para cada consulta diferente feita ao programa, *Calc_OI* ele deverá retornar um determinado valor de acordo com os valores de entrada.

Na tabela 4.5 especifica-se os tipos de dados propostos para a implementação, o fluxo de dados pode ser verificado na figura 4.4.

Inicialmente consideramos uma *função de vizinhança dependente do tempo* linear, que pode ser modificada para testes. O estudo de funções para problemas específicos faz parte das propostas para trabalhos futuros.

Capítulo 5

Testes e Resultados Obtidos

Os testes foram realizados com dados fornecidos pela CET/SP referentes à toda extensão da Rua da Consolação e alguns dos seus trechos. Na tabela 5.1 estes trechos estão listados.

Referem-se ao período de 01 a 30 de setembro de 2000, das 00 : 00hs às 23 : 30hs de cada dia. Retratam a velocidade média do tráfego a cada 15 minutos.

Os valores referentes à velocidade média foram obtidos por sensores situados nesta localidade. As informações referentes a cada trecho estão armazenadas em um total de onze arquivos. Cada arquivo contém informações referentes a um dos trechos. Para cada dia do mês temos velocidade média das 00 : 00hs às 23 : 30hs, a cada 15 minutos totalizando 90 valores de velocidade. Para um mês temos $90 \times 30 = 2700$ valores. Para os 11 arquivos temos 11×2700 totalizando 29700 valores de velocidade, aproximadamente, que foram disponibilizados para testes. A partir destes valores de velocidade, estimamos os valores do desvio padrão que também foram utilizados nos testes.

Esta é uma das localidades monitoradas pela CET devido ao grande fluxo de veículos e sua grande importância como via de tráfego na cidade. Localizada no centro de São Paulo, reflete com grande veracidade o retrato do tráfego nas principais vias públicas de uma grande metrópole.

Av. Paulista -Rua Fernando de Albuquerque
Rua Fernando de Albuquerque - Rua D. Antonieta de Queiroz
Rua D. Antonieta de Queiroz - Rua Caio Prado
Rua Caio Prado - AC. Ligação L/O
AC. Ligação L/O - Rua Nestor Pestana
Rua Nestor Pestana - Rua Amaral Gurgel
Rua Amaral Gurgel - Rua Maria Antonia
Rua Maria Antonia - Rua Sergipe
Rua Sergipe - Rua Maceió
Rua Maceió - Rua Antonio Carlos

Tabela 5.1: Tabela de trechos da rua da Consolação

5.1 Testes

Iniciamos nossos testes considerando uma perda de confiabilidade constante para todos os horários durante o dia, utilizando a função de vizinhança $f(t) = 0.01 \times t$. Primeiramente, testamos o nosso modelo, para os dados disponíveis, considerando a OI constante.

Para cada intervalo de tempo, neste caso a cada 15 minutos, a partir da velocidade média e desvio padrão obtivemos o gráfico de distribuição de probabilidade (ver figura 5.1) para o instante de recebimento t_0 .

Nesta figura observa-se o gráfico da distribuição normal de probabilidade para as diferentes classes de velocidade no instante do recebimento das informações.

Como próximo passo, precisávamos estimar a perda de confiabilidade desta distribuição com o decorrer do tempo. Assim, aplicamos o nosso modelo sobre os valores de probabilidades dados por esta curva obtemos duas novas curvas que representam os limitantes inferiores e superiores da variação de probabilidade com o tempo. Ver figura 5.2.

Nesta figura observa-se o gráfico da distribuição normal e duas curvas que representam os limitantes superiores e inferiores de OI.

A distância entre as curvas limitantes e a curva de probabilidade é dado por Ep que quantifica e representa a perda de confiabilidade com o tempo,

Figura 5.1: Gráfico de distribuição normal de probabilidade

Figura 5.2: Curvas limitantes superior e inferior de probabilidade

Figura 5.3: Gráfico de distribuição normal de probabilidade para $t_1 > t_0$

ou seja, a obsolescência de informação.

Desta forma nós obtivemos estimativas, da probabilidade inferior(P_i) e probabilidade superior(P_s) correspondentes aos pontos das curvas de probabilidade limitantes, respectivamente, inferior e superior, para cada classe de velocidade.

Mas precisávamos de um parâmetro para sabermos se estas estimativas, obtidas utilizando o nosso modelo, estavam corretas.

Visto que possuíamos os dados reais, disponíveis, podíamos obter os valores de probabilidade reais após o intervalo de tempo considerado, que é dado pela curva de distribuição a partir dos dados de velocidade média e desvio padrão para o instante $t_1 > t_0$. Ver figura 5.3.

Nesta figura observa-se o gráfico da distribuição normal de probabilidade para o instante $t_1 > t_0$.

A segunda etapa consistiu em verificar se as curvas limitantes da figura 5.2, que correspondem às estimativas da OI feitas pelo nosso modelo a partir das informações recebidas no instante t_0 , também são limitantes para a curva da figura 5.3, que correspondem à distribuição de probabilidade real para o instante $t > t_0$. Se isto ocorresse, poderíamos verificar a consistência do nosso modelo, visto que as estimativas foram obtidas a partir da sua utilização.

Feito isto, observamos que a perda de confiabilidade não era a mesma em todos os horários. Decidimos observar o comportamento da OI em busca de

possíveis padrões.

Na terceira etapa analisamos o comportamento da perda de confiabilidade dos dados reais disponíveis, com o decorrer do tempo. Obtivemos os gráficos referentes à distribuição de probabilidade de cada informação a cada 15 minutos e calculamos o valor da distância entre as funções de distribuição de probabilidade, representado por Ep .

Os valores de Ep obtidos quantificavam a perda de confiabilidade real, possibilitando, assim, observarmos o seu comportamento.

A partir destes valores, para cada dia disponível, construímos gráficos referentes à variação de Ep que correspondem à variação da OI. Estes gráficos não se referem à velocidade, mas sim, ao comportamento a perda de confiabilidade com o tempo. As figuras 5.4 e 5.5 ilustram alguns destes gráficos.

Os diversos gráficos foram agrupados de diferentes formas: trechos, dias da semana, dias do mês, etc, para observarmos possíveis padrões de comportamento.

A partir destes agrupamentos e da análise dos gráficos para diferentes valores de velocidade, pudemos observar que em determinados períodos do dia a OI apresentava oscilações acentuadas, e em outros períodos, apresentava oscilações menores. Fenômeno que se repetia com certa similaridade nos diferentes agrupamentos. Aprofundando esta análise, pudemos classificar a variação de Ep em virtude do tempo, conforme figura 5.6.

Nesta figura observa-se, para cada período de tempo diferente, a variação de Ep :

- nos horários $[0 - 6)$, $[8 - 10)$ e $[15 - 18)$ temos $Ep < 0.2$;
- nos horários $[10 - 15)$ e $[18 - 22)$ temos $Ep < 0.45$;
- nos horários $[6 - 8)$ e $[22 - 24)$ temos $Ep < 0.7$.

Classificamos a variação do Ep conforme tabela 5.2.

O que significa esta classificação?

Significa que a OI aqui observada não pode ser considerada constante durante o passar do tempo, sendo ela, sensível à períodos de tempo diferentes. Por outro lado, observamos que a OI obedece a um padrão perceptível que nos permitiu classificá-la.

Figura 5.4: Gráfico de OI

Figura 5.5: Gráfico de OI

Figura 5.6: Gráfico da variação de E_p

sigla	intervalo	variação
OIp	< 0.20	pequena
OIm	< 0.45	média
OIg	< 0.70	grande

Tabela 5.2: Tabela de classificação da variação da OI

- A primeira classe, denominada OIp, engloba os períodos no qual a perda de confiabilidade, de modo geral, é menor que 0.2.
- A segunda classe, denominada OIm, engloba os períodos no qual a perda de confiabilidade, de modo geral, é menor que 0.45.
- A terceira classe, denominada OIg, engloba os períodos no qual a perda de confiabilidade, de modo geral, é menor que 0.7.

A variação de probabilidade observada nos períodos referentes à primeira e terceira classes são, de modo geral, estáveis, obedecendo com muita fidelidade os limitantes de variação observados. No entanto, a variação de probabilidade observada no período pertencente à segunda classe, não é tão estável, apresentando, em alguns casos, variações que não seguem com fidelidade os limitantes desta classe, não invalidando, porém, a classificação.

Acreditamos que este comportamento é resultante de outros fatores, além do tempo. Fatores externos e particulares de cada trecho, como por exemplo colégios, comércio específico, etc, influenciam na velocidade e densidade do tráfego e conseqüentemente na OI.

A construção de um shopping, por exemplo, em um determinado trecho, irá alterar sensivelmente os resultados da OI.

Assim percebemos que a OI é sensível a fatores externos, relacionados ao problema específico, e que podem ser objeto de futuros estudos, mas que neste caso específico, está mais relacionado à engenharia de tráfego e urbanização do que à ciências da computação, no qual estamos nos concentrando.

A partir da classificação de *Ep*, reagrupamos os diversos gráficos e observando-os, percebemos algumas similaridades que nos permitiu classificá-los em três tipos.

- O primeiro tipo obedece fielmente à classificação da OI, o qual denominamos *Típico* (ver figuras 5.7 e 5.8).

Na figura 5.7 observa-se a variação da confiabilidade da informação com o tempo, correspondente ao trecho da Rua da Consolação entre Av. Paulista e Rua Fernando de Albuquerque, segunda feira, dia 04 de setembro de 2000.

Na figura 5.8 observa-se a variação da confiabilidade da informação com o tempo, correspondente ao trecho da Rua da Consolação entre

Rua Maceió e Rua Antonio Carlos, segunda feira, dia 25 de setembro de 2000.

- O segundo tipo obedece à classificação da OI, mas apresenta um comportamento diferenciado dentro das classes, normalmente observados em feriados e alguns finais de semana, o qual denominamos *Atípico* (ver figuras 5.9 e 5.10).

Na figura 5.9 observa-se a variação da confiabilidade da informação com o tempo, correspondente ao trecho da Rua da Consolação entre Av. Paulista e Rua Fernando de Albuquerque, quinta feira, dia 07 de setembro de 2000.

Na figura 5.10 observa-se a variação da confiabilidade da informação com o tempo, correspondente ao trecho da Rua da Consolação entre Av. Paulista e Rua Fernando de Albuquerque, domingo, dia 17 de setembro de 2000.

- O terceiro tipo apresenta grandes oscilações, principalmente na segunda classe de OI, o qual denominamos *Oscilante* (ver figura 5.11).

Nesta figura observa-se a variação da confiabilidade da informação com o tempo, correspondente ao trecho da Rua da Consolação entre Rua Sergipe e Rua Maceió, terça feira, dia 12 de setembro de 2000.

Agora, com esta nova análise e classificação, retomamos os testes realizados na primeira e segunda etapas, utilizando valores de Ep diferentes para cada classe de OI e obtivemos os resultados que descrevemos na próxima seção.

Figura 5.7: Gráfico de OI típico

Figura 5.8: Gráfico de OI típico

Figura 5.9: Gráfico de OI Atípico

Figura 5.10: Gráfico de OI Atípico

Figura 5.11: Gráfico de OI Oscilante

OI	Ep
OIp	0.10
OIm	0.25
OIg	0.35

Tabela 5.3: Tabela dos valores de Ep utilizados para testes

gráfico	acerto(%)
típico	90
atípico	80 – 90
oscilante	60 – 80

Tabela 5.4: Tabela de percentual de acertos para os diferentes tipos de gráficos

5.2 Resultados

A partir dos testes realizados efetuamos a modificação na função de vizinhança, que nos retorna o valor de Ep , de forma que ele seja sensível aos períodos da classificação da OI . No caso de um intervalo de tempo que englobe mais de uma classe, os valores de Ep devem ser acumulados.

A classificação dos valores de Ep para as classes OIp, OIm e OIg levou em conta o limitante superior para a perda de confiabilidade. Para nossos testes com o modelo reformulado, utilizamos os valores de Ep conforme tabela 5.3.

Esta escolha deve-se ao fato de que em cada classe, os valores máximos de obsolescência ocorrem muito pouco, sendo que os valores das oscilações, em média, correspondem ao valor aproximado da média de cada uma das classes. Desta forma não sub-estimamos nem super-estimamos os resultados da OI .

Refazendo os testes com os dados disponíveis, medimos o percentual de acertos na utilização do nosso modelo, agora reformulado para os valores de Ep , e obtivemos os resultados listados na tabela 5.4, que consideramos satisfatórios, para os diversos tipos de gráficos.

Esta porcentagem expressa a veracidade das respostas obtidas através da

utilização do nosso modelo, ou seja, na estimativa de variação de probabilidade obtivemos a porcentagem listada acima de respostas corretas.

Estamos lidando com informações imperfeitas (ver ex [17]), que perdem a confiabilidade com o tempo e que são sensíveis à fatores externos e muitas vezes não previsíveis. Diante deste fato, dificilmente podemos contar com acertos de 100%.

É importante observar que a nossa análise se restringiu à intervalos curtos de tempo. Uma análise mais profunda para intervalos de tempo maiores, como por exemplo 24 horas, permitirá novas investigações, como em particular a sazonalidade de certas informações. Para uma aplicação específica de tráfego, por exemplo, espera-se uma maior correlação entre 8 horas da manhã de um dia e 8 horas da manhã do dia seguinte do que entre um curto intervalo de tempo de um mesmo dia. O estudo aprofundado deste aspecto é proposto no capítulo 6 como proposta para trabalhos futuros.

Observamos que a hipótese da OI ser monotonicamente não decrescente com o tempo é válida para curtos intervalos de tempo e que obtivemos resultados satisfatórios com a aplicação do modelo proposto. Desta forma, concluímos que a obsolescência de informação pode ser modelada e o modelo proposto aplicado.

Capítulo 6

Discussão e Trabalhos Futuros

A pesquisa sobre a perda de confiabilidade da informação com o tempo é muito interessante pois pode ser encontrada em problemas de naturezas diferentes, e envolve várias áreas de conhecimento.

No nosso trabalho específico, além de inteligência artificial, envolveu engenharia de tráfego e urbanização, estatística e álgebra.

Estudamos e utilizamos o necessário destas áreas para nos concentrarmos no nosso foco de interesse, ciências da computação, e obtivemos resultados satisfatórios.

Partimos de um problema real e muito amplo- *como representar a perda de confiabilidade da informação no decorrer do tempo*. A partir deste problema verificamos os diversos estudos e lógicas conhecidas e utilizadas no tratamento de informações imperfeitas (ver ex. [2, 4, 6, 7, 8, 17]). Após este estudo, propusemos um modelo matemático genérico utilizando lógica probabilística e conceitos de distância e vizinhança entre funções de distribuição de probabilidade para tentarmos descrever o fenômeno da obsolescência de informação. Para a verificação da viabilidade do nosso modelo, utilizamos para testes dados referentes à velocidade média do tráfego, gentilmente fornecida pela CET/SP.

Através dos testes e resultados observados com a aplicação do modelo proposto, concluímos que a obsolescência de informação apresenta um comportamento perceptível e pode ser modelada. Ela apresenta variações diferentes para períodos de tempo que puderam ser padronizados. Ela é sensível

a fatores externos e o nosso modelo pode ser aplicado em sistemas que trabalhem com informações imperfeitas que perdem a confiabilidade com o tempo, sendo que ele deve ser adaptado às características particulares do problema a ser tratado.

Este trabalho representa o início de novas pesquisas de OI, que é muito ampla e nos oferece, ainda, vários problemas a serem solucionados.

Como proposta para futuros trabalhos apresentamos:

- Estudo da função de vizinhança dependente do tempo que é definida como alguma função real, não decrescente. Funções apropriadas podem ser selecionadas e calibradas a partir da avaliação dos dados de cada problema a ser modelado.
- Estudo estatístico do histórico dos dados e da OI do problema a ser modelado podem ser pesquisados para o detalhamento do comportamento da OI. Em particular informações podem ter sazonalidade que poderão ser observadas durante a análise, considerando intervalos longos de tempo.
- Estudo de fatores externos que influenciam a confiabilidade da informação. Nós observamos que alguns trechos específicos apresentam oscilações que não seguem com muita fidelidade ao padrão normal de OI observada. Isto se deve a fatores externos que podem ser pesquisados e aprofundados.
- Estudo de propostas de outros modelos matemáticos envolvendo a utilização de outras lógicas que também possam descrever o comportamento da OI.
- Revisão dos estudos realizados utilizando variáveis mais representativas da informação a ser analisada, como por exemplo o desvio padrão de velocidades típicas, apesar de ser mais difícil a obtenção dos dados relativos a estas variáveis.

Apêndice

A seguir, anexamos uma cópia de nosso artigo que foi apresentado na *International Joint Conferencia IBERAMIA-SBIA 2000* realizada em novembro de 2000 e publicado em seus anais no módulo *Open Discussion Track Proceedings*.

International Joint Conference
7th Iberoamerican Conference on Artificial Intelligence
15th Brazilian Conference on Artificial Intelligence
IBERAMIA-SBIA 2000

November 19-22, 2000
Atibaia, São Paulo, Brazil

Open Discussion Track
Proceedings

Maria Carolina Monard
Jaime Simão Sichman
(editors)

Sirley

**International Joint Conference
IBERAMIA-SBIA 2000**

**7th Iberoamerican Conference on Artificial Intelligence
15th Brazilian Conference on Artificial Intelligence**

**November 19-22 2000
Atibaia, São Paulo, Brazil**

**IBERAMIA - SBIA 2000
Open Discussion Track
Proceedings**

**Maria Carolina Monard
Jaime Simão Sichman
(editors)**

**ICMC/USP
São Carlos 2000**

Catalogação elaborada pela
Biblioteca Prof. Achille Bassi - ICMC/USP

I12i Iberoamerican Conference on Artificial Intelligence (7.: 2000 : Atibaia, SP, Brasil)
2000 Brazilian Conference on Artificial Intelligence (15.: 2000 : Atibaia, SP, Brasil)
IBERAMIA-SBIA 2000 open discussion track : proceedings / editors Maria Carolina Monard, Jaime Simão Sichman. - São Carlos, SP, Brasil : ICMC / USP, 2000.
x; 390 p.
ISBN 85-87837-03-6

1. Inteligência artificial - congressos. I. Monard, Maria Carolina, ed. II. Sichman, Jaime Simão, ed. III. Brazilian Conference on Artificial Intelligence (15.: 2000 : Atibaia, SP, Brasil). IV. Título.

AMS-68-06

Preface

This year, Brazil celebrates its 500 years of discovery. In order to mark this great event, the Brazilian Artificial Intelligence (AI) community organized a special international joint conference putting together SBIA 2000 (the Brazilian AI Symposium) and IBERAMIA 2000 (the Ibero-American AI Conference).

SBIA 2000 is the 15th conference of the SBIA conference series, which is the leading conference in Brazil for presentation of research and applications in Artificial Intelligence. Since 1995, SBIA has become an international conference, with papers written in English, an international program committee, and proceedings published in Springer-Verlag's *Lecture Notes in Artificial Intelligence* (LNAI) series.

IBERAMIA 2000 is the 7th conference of the IBERAMIA conference series, which has been one of the most suitable forums for ibero-american AI researchers (from South and Central America, Mexico, Spain, and Portugal) present their results. Following the SBIA and EPIA (Portuguese conference on AI) experiences, from IBERAMIA 98 on, it has also become an international conference, with proceedings published in Springer-Verlag's LNAI series.

The IBERAMIA-SBIA 2000 was held in Atibaia, São Paulo, Brazil, 19-22 November 2000. It consisted of a main conference, best MSc dissertation and PhD thesis contest, invited talks, some introductory and advanced tutorials and several workshops covering specific topics. The main conference was structured along two main modules: paper track and open discussion track.

The call for papers was very successful resulting in 156 submissions for paper track and 69 for open discussion track, from 18 different countries. In order to maintain the good standard of the conference, each submission was sent to at least three members of the program committee. A total of 48 papers were accepted for paper track presentation and were published in LNAI 1952. Another 33 papers were accepted for open discussion track and are included in this volume as well as the awarded PhD thesis and MSc dissertations from the 2nd Ibero-American contest chaired by Solange Oliveira Rezende.

We would like to thank all researchers for submitting their papers, the PC members and additional referees for the work they have done and all the sponsors who made the organization of IBERAMIA-SBIA 2000 possible. We are also very grateful to our colleagues who provided invaluable organizational support. The Scientific Chair would like to extend special thanks to Marcos R. Geromini and Gustavo E.A.P.A. Batista from ICMC-USP for their help and support while organizing the proceedings.

November 2000
São Paulo, Brazil

Maria Carolina Monard
Jaime Simão Sichman

Prefácio

Este ano, o Brasil celebra os seus 500 anos de descobrimento. Para marcar esse grande evento, a comunidade brasileira de Inteligência Artificial (IA) organizou uma conferência especial conjunta, integrando o SBIA 2000 (o Simpósio Brasileiro de IA) e o IBERAMIA 2000 (a Conferência Ibero-Americana de IA).

O SBIA 2000 é a 15ª da série de conferências do SBIA, a principal conferência no Brasil para apresentação de pesquisas e aplicações em Inteligência Artificial. Desde 1995 o SBIA é uma conferência internacional, com artigos escritos em inglês, um comitê de programa internacional e anais publicados na série *Lecture Notes in Artificial Intelligence* (LNAI) da Springer-Verlag.

O IBERAMIA 2000 é a 7ª da série de conferências do IBERAMIA, que tem sido um dos fóruns mais adequados para os pesquisadores ibero-americanos em IA (das Américas do Sul e Central, México, Espanha e Portugal) apresentarem os seus trabalhos. Seguindo as experiências do SBIA e do EPIA (conferência portuguesa em IA), a partir de 1998 o IBERAMIA se tornou uma conferência internacional com anais publicados na série LNAI da Springer Verlag.

O IBERAMIA-SBIA 2000 foi realizado em Atibaia, São Paulo, Brasil, de 19 a 22 de Novembro de 2000. O congresso consistiu de uma conferência principal, concursos de melhor dissertação de mestrado e de tese de doutorado, palestras convidadas, alguns tutoriais introdutórios e avançados e diversos *workshops* em tópicos específicos. A conferência principal foi estruturada em dois módulos principais: *Paper Track* e *Open Discussion Track*.

A chamada de trabalhos foi muito bem sucedida resultando em 156 submissões para o *Paper Track* e 69 submissões para o *Open Discussion Track*, de 18 países diferentes. Com o objetivo de manter o bom padrão da conferência, cada submissão foi enviada para pelo menos três membros do comitê de programa. Um total de 58 artigos foram aceitos para apresentação no *Paper Track*, sendo publicados na LNAI 1952. Outros 33 artigos foram aceitos para o *Open Discussion Track* e incluídos neste volume juntamente com as teses e dissertações premiadas no 2º concurso Ibero-Americano coordenado por Solange Oliveira Rezende.

Gostaríamos de agradecer a todos os pesquisadores por submeter os seus artigos, aos membros do comitê de programa e aos revisores adicionais pelo trabalho realizado, e a todos os patrocinadores que fizeram com que a organização do IBERAMIA-SBIA 2000 fosse possível. Estamos também muito agradecidos pelo trabalho de suporte à organização realizado por alguns de nossos colegas. A Coordenadora Científica gostaria de estender os seus agradecimentos especiais a Marcos R. Geromini e Gustavo E.A.P.A. Batista, do ICMC-USP, pela ajuda e suporte na organização dos anais.

Novembro de 2000
São Paulo, Brazil

Maria Carolina Monard
Jaime Simão Sichman

Organization

IBERAMIA-SBIA 2000 was organized by several AI research groups that belong to the University of São Paulo (USP):

D D D

Laboratório de Técnicas Inteligentes (LTI)
Departamento de Engenharia de Computação e Sistemas Digitais (PCS)
Escola Politécnica (EP)

D

Laboratório de Inteligência Computacional (LABIC)
Departamento de Ciências de Computação e Estatística (SCE)
Instituto de Ciências Matemáticas e de Computação (ICMC)

Laboratório de Lógica, Inteligência Artificial e Métodos Formais (LIAMF)
Departamento de Ciência da Computação (MAC)
Instituto de Matemática e Estatística (IME)

D

Organizing Committee

General Chair:	Jaime Simão Sichman (LTI/PCS/EP)
Scientific Chair:	Maria Carolina Monard (LABIC/SCE/ICMC)
Publicity Chair:	Fabio Nauras Akhras (LTI/PCS/EP)
Publication Chair:	Maria Carolina Monard (LABIC/SCE/ICMC)
Workshop Chair:	Leliane Nunes de Barros (LIAMF/MAC/IME)
Tutorial Chair:	Anna Helena Reali Costa (LTI/PCS/EP)
Finance Chair:	Edson Satoshi Gomi (LTI/PCS/EP)
Local Arrangements Chair:	Marco Túlio C. de Andrade (LTI/PCS/EP)

Supporting Scientific Societies

SBC	Sociedade Brasileira de Computação
AAIA	Asociación Argentina de Inteligencia Artificial
AEPIA	Asociación Española para Inteligencia Artificial
APPIA	Associação Portuguesa para Inteligência Artificial
AVINTA	Asociación Venezolana de Inteligencia Artificial
SCCC	Sociedad Chilena de Ciencia de la Computación
SMCC	Sociedad de Matemática y Computación de Cuba
SMIA	Sociedad Mexicana de Inteligencia Artificial

Table of Contents

Knowledge Engineering and Case Based Reasoning

Usando Conjuntos Borrosos para Implementar un Modelo para Sistemas Basados en Casos Interpretativos	1
<i>María Matilde García Lorenzo, Yanet Rodríguez Sarabia, Rafael Esteban Bello Pérez</i>	

User Profiling with Case-Based Reasoning and Bayesian Networks	12
<i>Silvia N. Schiaffino, Analía Amandi</i>	

Determinación y Manejo de la Incertidumbre en los Sistemas Basados en Casos	22
<i>I. Gutiérrez, R. Bello</i>	

Planning and Scheduling

Elaboração Automática de Horários de Aulas Combinando Técnicas de Programação Lógica com Restrições Hierárquicas e Simulated Annealing ..	37
<i>Wanderley Alencar, Cássio Vinhal</i>	

Distributed AI and Multi-Agent Systems

SACI: Uma Ferramenta para Implementação e Monitoração da Comunicação entre Agentes	47
<i>Jorní Fred Hübner, Jaime Simão Sichman</i>	

Coordination in MultiAgent Systems by Means of Linear Logic	57
<i>José Juan Palacios Pérez</i>	

Um Modelo Multi-Agente Descentralizado em Ambientes Fechados Baseado em Leis Sociais	67
<i>José Avelino Placca, Ana Cristina Bicharra Garcia</i>	

Autonomy through Value-driven Goal Adoption	77
<i>Luis Antunes, João Faria, Helder Coelho</i>	

Knowledge Representation and Reasoning

Remarks on Reasoning with Progressively Obsolete Information	87
<i>Flávio Soares Corrêa da Silva, Sirley Ambrósia Vitorio</i>	

Uso de Relacionamentos Semânticos para Auxiliar o Reconhecimento da Informação Manuscrita em Formulários	96
<i>Silvana Rossetto, Flávio M. Varejão, Thomas W. Rauber</i>	

Remarks on Reasoning with Progressively Obsolete Information

Flávio Soares Corrêa da Silva
Sirley Ambrósia Vitorio

Instituto de Matemática e Estatística, Universidade de São Paulo
Rua do Matão, 1010. São Paulo - SP, Brasil. 05508-900.
E-mail: fcs@ime.usp.br, sirley@ime.usp.br

Abstract

Reasoning with uncertain information is a central issue for Artificial Intelligence and Information Systems development. Consequently, a vast amount of studies and results devoted to this issue can be found in the literature.

Curiously, reasoning with information whose uncertainty *varies* along time cannot be found so much in the literature. Dynamic reliability is a feature found rather frequently in many problems, as in e.g. finance and economics, sociology, and demographics.

In the present article we discuss this issue and propose a model for information that becomes progressively unreliable with time. Our motivating case study is the information system under development within the project *SIDAM*¹, to provide information about urban automotive traffic for large cities.

¹Distributed Information Systems for Mobile Agents – supported by FAPESP budget nr. 98/06138-2.

1 Introduction

Reasoning with uncertain information is a central issue for Artificial Intelligence and Information Systems development. Consequently, a vast amount of studies and results devoted to this issue can be found in the literature (see e.g. [Bac90, Bux89, Cla90, CdS99, CdSRH94]).

Curiously, reasoning with information whose uncertainty *varies* along time cannot be found so much in the literature. Dynamic reliability is a feature found rather frequently in many problems. For example:

1. finance and economics: stocks markets analysts rely upon information that must be continuously updated, lest the efficiency of their suggested operations decreases. Given the characteristic competitiveness of these markets, failure to update information for some analysts is regarded as an investment opportunity for the others. Hence, market opportunities in stock exchange are generated precisely by the dynamics of reliability of information available to different analysts;
2. sociology and demographics: measured behaviour of populations (migration rates, city growth rates, social stability rates, unemployment rates, population density and activities distribution, etc.) are clearly dynamical. These measures are also expensive to be produced, and hence they can only be updated periodically. Between measures, estimates must be made (typically via extrapolation of values based on time series analyses). These estimates become less reliable the further we go from actual measures. The (progressively decreasing) reliability of estimated information is what determines the frequency of measures to be performed.

In the present article we discuss this issue and propose a model for information that becomes progressively unreliable with time. Our motivating case study is the information system under development within the project *SIDAM - Distributed Information Systems for Mobile Agents* [EdSeS⁺00, FCdS98]. The goal of the project *SIDAM* is the provision of technical infrastructure for distributed information systems maintained and queried by mobile agents. A typical scenario for application of this infrastructure is an information system about urban vehicle traffic in big cities, in which users are drivers interested in making decisions dynamically about their routes, to avoid traffic jams.

We are developing the project based on data available about the city of São Paulo (BRAZIL), kindly provided to this research initiative by the local Traffic Engineering Company (CET/SP). The information is obtained from underpavement loop sensors, that allow estimates of traffic density and average speed at fixed intervals of fifteen minutes.

We assume that drivers will select their routes based on expected traffic density and average speed at different points in the city. Different reliability degrees for information appear in this setting for two reasons:

1. information feed occurs at every two minutes for each sensor (typically located at representative road junctions), but not necessarily synchronously.

Moreover, since "pointwise" information obtained at every two minutes is not so informative, CET/SP has adopted the procedure of accumulating information received at every fifteen minutes, to produce statistical estimates of values of variables in that period; and

2. typically a driver will be interested not on the present traffic status of different junctions, but on the forecasted status for when each junction shall be reached.

We assume that we have, for each junction, a fresh estimate for the probabilities of all possible traffic densities and average speeds coming every fifteen minutes. In the following sections we present how the dynamic loss of reliability on these estimated probabilities can be modeled and implemented. In section 2 we introduce our proposed model and show how it is expected to be used within the context of *SIDAM*. In section 3 we present a computational experiment for this model. In section 4 we detail the implementation used in this experiment. Finally, in section 5 we present some further discussion and proposed future work.

2 A Model for the Dynamics of Obsolescence of Information

For the purposes of the intended application of this work, we assume that information has been "discretised", i.e. that observed values for traffic density and average speed have been classified to constitute a finite domain of "relevant values".

Each of these values will be called an *atomic event*. What the sensors will provide at every fifteen minutes are *atomic probability estimates* for the finite set of atomic events.

The set of atomic events is denoted as $E = \{e_1, \dots, e_n\}$. An atomic probability estimate is a real function $\mu_e : E \rightarrow [0, 1]$, such that $\sum_E \mu_e(e_i) = 1$.

An *event* is any subset of E . The *probability estimate* for the set of events based on the atomic probability estimate μ_e is the real function $\mu : 2^E \rightarrow [0, 1]$ such that $\mu(\mathcal{E}) = \sum_{e_i \in \mathcal{E}} \mu_e(e_i)$, $\mathcal{E} \in 2^E$.

Given two atomic probability estimates $\mu_{e1}, \mu_{e2} : E \rightarrow [0, 1]$, we define the *distance* between the corresponding probability estimates as $D(\mu_1, \mu_2) = \max\{|\mu_{e1}(e_i) - \mu_{e2}(e_i)| : e_i \in E\}$. We define the δ -neighbourhood of a probability estimate $\mu : 2^E \rightarrow [0, 1]$ ($\delta(\mu)$) as:

$$\delta(\mu) = \{(\mu' : 2^E \rightarrow [0, 1]) : D(\mu, \mu') \leq \delta\}$$

Clearly, for $\delta \geq 1$, the δ -neighbourhood of any probability estimate becomes the same, i.e. the set of all possible probability estimates for a given set of events 2^E .

Now we are ready to introduce our proposed model for the obsolescence of information. Given a linear, real-valued time axis $T = [0, \infty)$, and assuming that

Figure 1: Obsolence of Information: Fixed Time Period

we reset the time to 0 whenever information is updated by the sensors, we define a *time-dependent neighbourhood function* $\delta : T \rightarrow [0, 1]$ as any non-decreasing function such that $\delta(0) = 0$.

Given a probability estimate μ_0 for $t_0 = 0, t \in T$, we simply assume that the probability estimate for $t > 0$ will belong to the δ -neighbourhood of μ_0 , where $\delta = \delta(t)$. As a consequence, given an event $\mathcal{E}$, we cannot access the probability estimate $\mu(\mathcal{E})$ when $t > 0$, because we do not know what probability estimate is the actual one – all we know is that $\mu \in \delta(t)(\mu_0)$. All we can say about the probability estimate for $\mathcal{E}$ is that, for each atomic event $e \in \mathcal{E}$, its corresponding atomic probability estimate will belong to the interval

$$[\max\{0, \mu_0(e) - \delta(t)\}, \min\{1, \mu_0(e) + \delta(t)\}].$$

Given an event of interest $\mathcal{E}$ and a time $t \geq 0$, we can be interested in determining the interval to which the probability estimate $\mu(\mathcal{E})$ is bound to belong. Alternatively, given an event of interest and a probability value, we can be interested in determining how long it will take for that value to become liable to be the real one.

In figures 1 and 2 we illustrate these two possibilities. In figure 1 we depict the dynamics of the intervals for probability estimates, showing the expected evolution of an interval after a fixed time period. In figure 2 we depict the same dynamics of the intervals for probability estimates, showing the time period needed for a fixed probability value to become liable to be the real one.

3 Experimenting with this Model

For the purposes of the intended application of this work, we assume that probabilistic information gathered by the underpavement loop sensors (i.e. traffic

Figure 2: Obsolescence of Information: Fixed Probability Value

speed) follows a *Normal Distribution Model* [Men85]. At every fifteen minutes, the sensors generate new estimates for *expected speed* and *standard deviation* (σ) of traffic speed.

Information is "discretised" as a finite amount of intervals of fixed length. In our case, traffic speed (given in *km/hour*) is "discretised" in ten speed classes, as presented below:

class	speed interval
1	[0, 20)
2	[20, 40)
3	[40, 60)
$\vdots$	$\vdots$
10	[180, 200)

In the assumed probabilistic model, the probability that a continuous random variable assumes a value in an interval $[a, b]$ is equals the area under its probability density function between points a and b [Men85]. Hence, given the expected speed and standard deviation as above, we can estimate the probability of each class corresponding to a speed interval.

Given a time-dependent neighbourhood function $\delta(t)$, defined as a specific non-decreasing function according to the problem under consideration, we can estimate a probability interval for each class at any time period t after a given probabilistic estimate generated by processing information given by the loop sensors.

In our prototypical implementation, we assume as external data given to the program the following:

1. expected speed avg ;

2. standard deviation σ ;
3. atomic event v ; and
4. time period $t \geq 0$.

The program then finds to which *class* cl the atomic event v belongs, then estimates the probability $\mu_0(cl)$ of that class based on avs and σ , and finally estimates the probability interval determined by $\delta(t)$ as:

$$[\max\{0, \mu_0(cl) - \delta(t)\}, \min\{1, \mu_0(cl) + \delta(t)\}].$$

Alternatively, we can be interested in solving the following problem: given a speed class and a probability value, determine how long it will take for that value to become liable to be the real one.

In this case, we assume as external data given to the program the following:

1. expected speed avs ;
2. standard deviation σ ;
3. atomic event v ; and
4. probability value μ .

The program then finds (as before) to which *class* cl the atomic event v belongs, then estimates the probability $\mu_0(cl)$ of that class based on avs and σ and determines the minimum t such that:

$$\mu \in [\max\{0, \mu_0(cl) - \delta(t)\}, \min\{1, \mu_0(cl) + \delta(t)\}].$$

Initially, we have implemented a linear time-dependent neighbourhood function. More appropriate functions for our specific problem shall be topic of future research.

4 Implementation

The program has been implemented in standard ANSI C. It is composed of one set of functions fixed: *Calc_Class*, *Calc_Prob*, *Calc_InterProb*, *Calc_Func*, *Calc_Temp*, *In_Cet*. These functions perform our mathematical model.

The function *Calc_OI* apply the functions from the model through several ways.

For every different queries made by the user, *Calc_OI* return a value according to the values input.

Example: *Calc_OI1* reads speed and period of time and returns interval probability $[Pi, Ps]$.

Calc_OI2 reads speed and value of probability and returns period of time which this probability is valid.

The program is composed by the following functions (see figure 3):

Figure 3: Implementation: information flow

```

void In_Cet(float *avs, float *sd, hour *H0);
int Calc_Class(float v, int sz);
double Calc_Prob(int cl, int sz, float avs, float sd);
double Calc_Func(float t);
interp Calc_InterProb(double P0, double Epi);
float Calc_Time(double P0, double PE);
interp Calc_OI1(float v, hour He);
hour Calc_OI2(float v, double PE);

```

Ps: *interp* is one structure composed by two values $\text{double}[P_{inf}, P_{sup}]$, and *hour* is one structure for hours

In_Cet: reads input values provided for *CET* – expected speed(*avs*), standard deviation(*sd*), time of input data (*H0*).

Calc_Class: given speed *v* and class size *sz* (in our case, class size is 20), returns to what class *cl* the given speed belongs.

Calc_Prob: given the class *cl*, class size *sz*, expected speed(*avs*) and standard deviation(*sd*) returns the estimated initial probability *P0* for time $t = t0$.

Calc_Func: given a time period *t*, returns the neighbourhood $\delta(t)$.

Calc_InterProb: given the estimated initial probability *P0* returned by *Calc_Prob* and the neighbourhood *Epi* returned by *Calc_Func*, returns the interval $[P_i, P_s]$.

Calc_Time: given the estimated initial probability *P0* returned by *Calc_Prob*, the target probability value *PE*, iteratively generates intervals using *InterProb* with different time periods until we have a time period *te* such that $E \in [P_i, P_s]$.

Calc_OI1: given speed *v* and one time *He*, returns the interval $[P_i, P_s]$.

Calc_OI2: given the speed and the value probability *PE* returns the time what the probability *PE* is valid.

As an illustration, given $avs = 60 \text{ km/hour}$, $\sigma = 5 (\text{km/hour})^2$ and $v = \text{km/hour}$, using the function $\delta(t) = 0.01 * t$, we obtain the following results

for different values of t :

time (t)	interval
0	$0.50 \leq \mu_0 \leq 0.50$
2	$0.48 \leq \mu_0 \leq 0.52$
5	$0.45 \leq \mu_0 \leq 0.55$
10	$0.40 \leq \mu_0 \leq 0.60$
20	$0.30 \leq \mu_0 \leq 0.70$
30	$0.20 \leq \mu_0 \leq 0.80$
40	$0.10 \leq \mu_0 \leq 0.90$
50	$0.00 \leq \mu_0 \leq 1.0$
60	$0.00 \leq \mu_0 \leq 1.0$

Given $av_s = 60\text{km/hour}$, $\sigma = 5(\text{km/hour})^2$ and $v = 70\text{km/hour}$, using the function $\delta(t) = 0.07 * t$, we obtain:

time (t)	interval
0	$0.50 \leq \mu_0 \leq 0.50$
2	$0.36 \leq \mu_0 \leq 0.64$
4	$0.22 \leq \mu_0 \leq 0.78$
6	$0.08 \leq \mu_0 \leq 0.92$
8	$0.00 \leq \mu_0 \leq 1.00$
10	$0.00 \leq \mu_0 \leq 1.00$

5 Discussion and Future Work

The framework presented above is very straightforward and easy to implement, and a prototype has been implemented by the authors as presented.

Two major research issues arise from a preliminary analysis of this model and how it can be used:

1. the time-dependent neighbourhood function was purposefully underdefined as any non-decreasing real-valued function. Appropriate functions shall be selected and calibrated depending on available data for each problem being modeled.
2. the information conveyed by this model can be exploited in many ways to aid decision-making. For example, one may be interested in, given discrete probability distributions and time-dependent neighbourhood functions for several variables, how to evaluate boolean combinations of these variables as regards the two informations above.

These topics are being studied at the moment and shall be subject of future articles.

Acknowledgements: This work is partially supported by FAPESP grant *SIDAM* - nr. 98/06138-2.

References

- [Bac90] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, 1990.
- [Bux89] R. Buxton. Modelling Uncertainty in Expert Systems. *International Journal of Man-Machine Studies*, 31:415-476, 1989.
- [CdS99] F. S. Correa da Silva. Formalisations of Uncertain Reasoning (extended abstract). In Springer-Verlag, editor, *12th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (Lecture Notes in Artificial Intelligence, v. 1611)*, pages 1-3, Cairo (Egypt), 1999. D
- [CdSRH94] F. S. Correa da Silva, D. S. Robertson, and J. Hesketh. Automated Reasoning with Uncertainties. In M. Masuch and L. Pawlos, editors, *Knowledge Representation and Uncertainty - ch. 5 - a preliminary version was presented at the APLOC - Applied Logic Conference (Logic at Work), 1992*. Springer Verlag LNAI 808, 1994.
- [Cla90] D. A. Clark. Numerical and Symbolic Approaches to Uncertainty Management in AI. *Artificial Intelligence Review*, 4:109-146, 1990.
- [EdSeS⁺00] M. Endler, D.M. da Silva, F. Silva e Silva, R.A. da Rocha, and M.A. de Moura. Project SIDAM: Overview and Preliminary Results. In *Anais do 2o. Workshop de Comunicação sem Fio (WCSF), Belo Horizonte*, May 2000. (submitted). D
- [FCdS98] M. Finger and F. S. Correa da Silva. Temporal Data Obsolescence: Modeling Problems. In *Proceedings of the 5th International Workshop on Temporal Representation and Reasoning*, USA, 1998.
- [Men85] W. Mendelhall. *Probabilidade e Estatística (portuguese translation)*. Campus, 1985.

Referências Bibliográficas

- [1] G. Ávila. *Cálculo 3*. Editora LTC, 1983.
- [2] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, 1990.
- [3] L. C. Barroso, M. M. A. Barroso, F. F. C. Filho, M. L. B. Carvalho, and M. L. Maia. *Cálculo Numérico*. Harper & Row do Brasil, 1983.
- [4] R. Buxton. Modelling Uncertainty in Expert Systems. *International Journal of Man-Machine Studies*, 31:415–476, 1989.
- [5] A. F. P. Castro Humes, I. S. Homem de Melo, L. K. Yoshida, and W. T. Martins. *Nocões de Cálculo Numérico*. Mc Graw Hill, 1984.
- [6] D. A. Clark. Numerical and Symbolic Approaches to Uncertainty Management in AI. *Artificial Intelligence Review*, 4:109–146, 1990.
- [7] F. S. Correa da Silva. Formalisations of Uncertain Reasoning (extended abstract). In Springer-Verlag, editor, *12th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (Lecture Notes in Artificial Intelligence, v. 1611)*, pages 1–3, Cairo (Egypt), 1999.
- [8] F. S. Correa da Silva, D. S. Robertson, and J. Hesketh. Automated Reasoning with Uncertainties. In M. Masuch and L. Pawlos, editors, *Knowledge Representation and Uncertainty – ch. 5 – a preliminary version was presented at the APLOC – Applied Logic Conference (Logic at Work), 1992*. Springer Verlag LNAI 808, 1994.

- [9] M. Endler, D.M. da Silva, F. Silva e Silva, R.A. da Rocha, and M.A. de Moura. Project SIDAM: Overview and Preliminary Results. In *Anais do 2o. Workshop de Comunicação sem Fio (WCSF)*, Belo Horizonte, May 2000. (submitted).
- [10] M. Finger and F. S. Correa da Silva. Temporal Data Obsolescence: Modeling Problems. In *Proceedings of the 5th International Workshop on Temporal Representation and Reasoning*, USA, 1998.
- [11] K. Hoffmann and Kunze R. *Álgebra Linear*. Universidade de São Paulo, 1971.
- [12] W. Kaplan and D. J. Lewis. *Cálculo e Álgebra Linear*. LTC, 1977.
- [13] E. L. Lima. *Espacos Métricos*. Impa, CNPq, 1977.
- [14] W. Mendelhall. *Probabilidade e Estatística*. Campus, 1985.
- [15] G. Milone and F Angelini. *Estatística Geral*. Editora Atlas, 1993.
- [16] S. H. Ming. *Uma Breve Descrição do Sistema SCOOT*. Séries Novas Técnicas. CET- São Paulo nr 201, 1997.
- [17] A. Motro and F. Smits. *Uncertainty Management in Information Systems*. Kluwer Academic, 1997.
- [18] L. Nachbin. *Introdução à Análise Funcional, Espacos de Banach e Cálculo Diferencial*. Editora LTC, 1984.
- [19] P. A Szasz. *Obtenção do Tempo Médio de Percurso a partir dos Dados do SCOOT/ITACA*. Relatório de Trabalho- CET São Paulo, 1998.
- [20] L. M. VilaNova. *Processo de Automatização do Cálculo das Velocidades-Infovel*. Relatório de Trabalho- CET São Paulo, 1998.