Mineração Adaptativa de Dados: Aplicação à Identificação de Indivíduos

P. R. M. Cereda e J. J. Neto

***Resumo*— Este artigo apresenta uma proposta de modelo adaptativo para identificação de indivíduos voltada a grandes volumes de dados. Adicionalmente, apresenta-se também uma proposta inicial do conceito de mineração adaptativa de dados como possível solução computacional aplicável aos problemas oriundos de grandes volumes de dados.**

***Palavras-chave:*— Dispositivos Adaptativos, Identificação de Indivíduos, Mineração de Dados.**

I. Introdução

Com o advento e popularização dos serviços web e das redes sociais, o volume de dados disponíveis aumentou exponencialmente nos últimos anos. Estudos recentes estimam os totais de usuários ativos cadastrados, e os números impressionam: a rede de relacionamento *Facebook* lidera o *ranking* de cadastros, possuindo atualmente 500 milhões de usuários, seguida pelo serviço de e-mail *GMail*, com 190 milhões, e pela rede *MySpace*, com 126 milhões [1]. As informações de tais redes e serviços, quando analisadas e interpretadas, podem oferecer subsídios para um mapeamento mais preciso sobre perfis de consumo e padrões comportamentais de seus usuários. Entretanto, o maior desafio é como atuar sobre grandes volumes de dados em tempo hábil.

Técnicas tradicionais de mineração de dados para extração de conhecimento são amplamente utilizadas em grandes volumes de dados, porém muitas delas apresentam tempo de execução relativamente lento ou ainda, no pior caso, de ordem exponencial [2]. Bancos de dados de serviços web e de redes sociais já ultrapassam milhares de *terabytes*, e a utilização de técnicas tradicionais pode exaurir facilmente recursos computacionais, demandar tempo impraticável e inviabilizar o processo de extração de conhecimento.

A *adaptatividade* é o termo utilizado para denotar a capacidade de um dispositivo em modificar seu próprio comportamento, sem a interferência de agentes externos. Existem inúmeras aplicações para a adaptatividade em sistemas computacionais, incluindo robótica [3], tomadas de decisão e aprendizado de máquina [4], [5], processamento de linguagem natural [6], otimização de código em compiladores [7], controle de acesso [8], servidores web [9], linguagens de programação [10], [11], meta-modelagem [12], [13] e computação evolutiva [14], [15]. De acordo com Neto [16], a generalidade resultante do modelo, a capacidade de aprendizado devida à característica de auto-modificação, bem como o poder de expressão faz dos dispositivos adaptativos uma alternativa muito atraente para expressar fatos complexos e manipular situações difíceis que surgem durante uma busca de soluções computacionais para problemas complexos.

Este artigo apresenta uma proposta de modelo adaptativo para identificação de indivíduos em meio a grandes populações. O autômato adaptativo foi escolhido como dispositivo do modelo devido à sua simplicidade de abstração e poder computacional. É importante destacar que outros dispositivos adaptativos poderiam ser utilizados neste mesmo contexto. Além do modelo, apresenta-se também uma proposta inicial do conceito de mineração adaptativa de dados como possível solução computacional aplicável aos problemas oriundos de grandes volumes de dados.

A organização deste artigo é a seguinte: a Seção II contextualiza a identificação de indivíduos, cálculo e aplicações. A mineração adaptativa de dados é apresentada na Seção III, juntamente com a mineração de dados tradicional, exemplos e aplicações. A Seção IV apresenta um experimento realizado para analisar o impacto do modelo adaptativo e algumas possíveis vantagens da proposta deste artigo. As discussões sobre o modelo e a mineração adaptativa de dados são apresentadas na Seção V. A Seção VI apresenta as conclusões.

II. Identificação de Indivíduos

A *identificação de indivíduos*, no contexto deste artigo, é definida como um método aplicado a um conjunto de indivíduos, reduzindo-o a um subconjunto com características específicas comuns. Se o subconjunto resultante possuir apenas um elemento, a identificação é dita *única*. No caso de um subconjunto vazio, a identificação sob as características comuns não é aplicável.

Do ponto de vista comercial, a identificação de indivíduos oferece subsídios para a criação de segmentos de mercado. Através de tais segmentos, é possível estabelecer, heuristicamente, perfis de consumo e padrões comportamentais. No caso de uma identificação única, o grau de corretude das informações relacionadas disponíveis é muito maior, mas pode incorrer em eventuais violações de privacidade [17], [18].

A aquisição das características dos indivíduos de um conjunto pode ocorrer de dois modos, independentes ou não

Os autores podem ser contatados através dos seguintes endereços de correio eletrônico: cereda@users.sf.net e joao.jose@poli.usp.br.

entre si: o primeiro, *explícito*, é oriundo de interações diretas de um indivíduo com os mais diversos tipos de coletores de dados, incluindo formulários de empresas e bancos, currículos, cadastros de prestações de serviços, entrevistas, concursos públicos. Espera-se que, neste caso, o grau de corretude das informações coletadas seja muito elevado, pois a manipulação e alteração propositais de tais informações por parte do indivíduo pode constituir crime de falsidade ideológica [19].

O segundo modo trata de uma aquisição *implícita*, sem necessariamente o conhecimento e consentimento do indivíduo. A aquisição ocorre em diversos níveis de interação indireta e resulta em um fluxo enorme de informações [20], [9]. Os coletores de dados mais comuns na aquisição implícita atuam no escopo da internet e incluem diversos mecanismos e técnicas disponíveis, tais como *crawlers*, *spiders* e *web bugs* [21], [22], [23], [24], [25]. Com o crescimento e difusão do comércio eletrônico e da popularização de redes sociais, a aquisição implícita tornou-se fundamental na tentativa de identificação de indivíduos com o máximo de características disponíveis.

Além das interações em serviços web, os possíveis padrões comportamentais podem ser também monitorados ou rastreados através de serviços baseados em localização. Tais serviços utilizam milhares de informações de campo, como coordenadas geográficas, temperatura externa, sons e ruídos, para definir o conjunto de dados a ser processado [26] [27] [28] [29] [30]. A computação móvel ainda permite que dispositivos em sistemas ubíquos e pervasivos se comuniquem e troquem dados entre si, o que aumenta expressivamente o conjunto de dados disponível acerca de um indivíduo e de seu comportamento [31].

Outra forma de rastreamento ocorre através de *impressões digitais*, em hardware e software. Na área de hardware, existem diversos registros na literatura sobre dispositivos que possuem características e variações únicas [32], [33], [34], [35]. Essas informações favorecem uma eventual identificação do dispositivo e consequente rastreamento de seu possuidor.

Na área de software, o rastreamento pode ou não ocorrer de modo deliberado. Alguns aplicativos utilizam métodos para geração e envio de *hashes* ou qualquer outra informação desejável a um determinado servidor. Esse procedimento é amplamente utilizado para verificação e validação de chaves de registros. Em outros casos, um aplicativo pode fornecer informações de acordo com uma determinada solicitação. Por exemplo, um navegador web envia sua própria identificação e versão juntamente com dados do sistema operacional no cabeçalho HTTP [36]. Existem estudos que analisam a impressão digital de navegadores web através das informações contidas no cabeçalho HTTP juntamente com a lista de *plugins* e complementos disponibilizados [37], e seus resultados são impressionantes: um em aproximadamente 287 mil navegadores compartilhará a mesma impressão digital que outro navegador, e em 99,1% dos casos é possível identificar um navegador reincidente, mesmo que este tenha sido atualizado.

Técnicas de proteção de identidade e melhoria de privacidade foram propostas para tentar atenuar a exposição de dados sensíveis [38], [39], [40], [41]. Algumas delas utilizam-se do conceito de *anonimato de grupo*, que consiste na situação em que um indivíduo se torna anônimo através do grupo em que ele está inserido. Um dos exemplos mais intuitivos do anonimato de grupo é o batedor de carteiras, que após praticar o delito, corre em direção à multidão ao invés de lugares mais abertos; quanto maior a multidão, mais difícil será encontrá-lo. Seguindo esta lógica, seria razoável imaginar que, dada a população mundial – aproximadamente 7 bilhões de indivíduos em 2011 [42] – o anonimato de um indivíduo, no âmbito global, estaria preservado. Entretanto, estudos indicam que cada indivíduo é facilmente rastreável através de análises das informações coletadas [37], [43].

Uma das métricas utilizadas para a análise de identificação de um indivíduo é o *cálculo dos bits de informação*, obtido através da Fórmula 1 [43].

$$\Gamma = -\log_2 P(A) \tag{1}$$

De acordo com a Fórmula 1, o valor obtido representa a quantidade de informação necessária para se identificar um indivíduo. É calculado utilizando-se o logaritmo binário de uma probabilidade *P(A)* dada [43]. Como exemplo, considere a probabilidade de se encontrar um indivíduo em 7 bilhões – a população mundial – como sendo $P(A) = 1/7000000000$. Assim, o cálculo de Γ_1 é demonstrado na Fórmula 2 [43].

$$\Gamma_1 = -\log_2 \frac{1}{7000000000} \tag{2}$$

O resultado obtido na Fórmula 2 é de 32,7047 indicando que apenas 32,7 bits de informação são suficientes para se identificar univocamente um indivíduo entre a população mundial. É importante observar que 32,7 bits de informação representa o valor obtido no pior caso, pois quanto menor o grupo, menor é o número de bits de informação necessário para se identificar univocamente um indivíduo nesse grupo.

O cálculo do número de bits de informação permite não somente a identificação única de um indivíduo, mas pode indicar uma classificação em subconjuntos contendo indivíduos com características semelhantes [43]. A partir desses grupos, perfis de consumo e padrões comportamentais são extraídos através das mais diversas técnicas de mineração de dados.

III. Mineração Adaptativa de Dados

Mineração de dados é o nome associado ao conjunto de técnicas para aquisição de conhecimento relevante em grandes volumes de dados. Também conhecida como *data mining*, a mineração de dados constitui a segunda fase em um processo de *descoberta de conhecimento em bases de dados*, um ramo da computação que utiliza ferramentas e técnicas computacionais para sistematizar o processo de extração de conhecimento [44].

A primeira fase do processo de descoberta de conhecimento é chamada de *preparação de dados*, na qual os dados são pré-processados para a posterior submissão ao

processo de mineração. É composta das seguintes etapas, a saber:

a) *definição dos objetivos do problema*, que constitui o entendimento do domínio da aplicação e quais os objetivos a serem alcançados,
b) *criação do conjunto de dados*, na qual o conteúdo bruto de dados a serem analisados é agrupado e organizado,

filtragem e pré-processamento de dados, tendo como objetivo assegurar a qualidade dos dados selecionados; esta etapa é uma das mais onerosas de todo o processo – pode atingir até 80% do tempo estimado – principalmente devido às dificuldades de integração de conjuntos de dados heterogêneos [45],

c) *redução e projeção de dados*, que constitui a alocação e armazenamento adequados em bancos de dados para facilitar a aplicação das técnicas de mineração de dados.

A segunda fase consiste na *mineração de dados* propriamente dita sobre os dados preparados na fase anterior. É composta das seguintes etapas, a saber:

a) *escolha das técnicas de mineração*, na qual selecionam-se uma ou mais técnicas para atuar sobre o conjunto de dados de acordo com o problema definido; não existe uma técnica universal, portanto cada problema exige uma escolha de quais técnicas são mais adequadas,
b) *mineração*, que efetivamente aplica as técnicas no conjunto de dados,
c) *interpretação dos padrões de exploração*, que faz a análise da informação extraída em relação aos objetivos definidos,
d) *consolidação do conhecimento descoberto*, que consiste na filtragem das informações, eliminando possíveis padrões redundantes ou irrelevantes, gerando assim o conhecimento a partir da fase de mineração de dados.

As técnicas de mineração de dados apresentam características e objetivos distintos, de acordo com a descrição de cada problema em particular. As mais tradicionais incluem classificação, relacionamento entre variáveis, agrupamento, sumarização, modelo de dependência, regras de associação e análise de séries temporais [46], todas elas utilizando conceitos das áreas de aprendizado de máquina, reconhecimento de padrões e estatística. Propõe-se neste artigo a inserção e utilização dos conceitos da área de tecnologia adaptativa no processo de descoberta de conhecimento em bancos de dados, possibilitando que a mineração de dados se torne adaptativa.

A *mineração adaptativa de dados* é definida como sendo um conjunto de técnicas adaptativas para aquisição de conhecimento. Através da utilização de dispositivos adaptativos, espera-se reduzir o número de etapas do processo de descoberta de conhecimento em bancos de dados e o número de passos computacionais. Além disso, a mineração adaptativa de dados, através de modelos adaptativos, pode permitir a análise, mineração e interpretação incremental dos dados, independentes do término das etapas de coleta, pré-processamento e verificação.

A tecnologia adaptativa tem obtido resultados expressivos na resolução de problemas referentes à extração de perfis de consumo e padrões comportamentais. Estudos recentes demonstraram que um dispositivo adaptativo pode, por exemplo, substituir com sucesso uma técnica de extração baseada em regras de associação muito utilizada no processo de mineração de dados [9]. As vantagens da utilização de um dispositivo adaptativo sobre as técnicas tradicionais incluem:

a) *tempo de execução linear*, mesmo para grandes conjuntos de dados,
b) *simplicidade formal e computacional do modelo*,
c) *capacidade de auto-modificação*, sem a necessidade de interferência externa de um especialista de domínio,
d) *modelo de dados disponível de modo incremental*,
e) *consistência contínua*, não havendo a necessidade de filtragem do conhecimento gerado [9].

O autômato adaptativo [47], [48] foi escolhido como dispositivo adaptativo a ser utilizado para a definição de um modelo de extração de conhecimento e um estudo de caso. Suas características permitem um modelo consistente, simplificado e poderoso computacionalmente [16], além de oferecer subsídios para uma implementação direta a partir da definição proposta.

O modelo apresentado a seguir tem como objetivo a extração de conhecimento a partir de um conjunto de dados disponível, na tentativa de identificação de indivíduos a partir de suas características. Inicialmente, o autômato adaptativo *M* é definido com uma configuração que reflete o conjunto de dados disponível (Figura 1). Durante seu tempo de vida, *M* sofre alterações em sua topologia para extrair conhecimento e identificar indivíduos. O caso aqui apresentado é um exemplo ilustrativo e pode ser adequado para ser utilizado em outras aplicações.

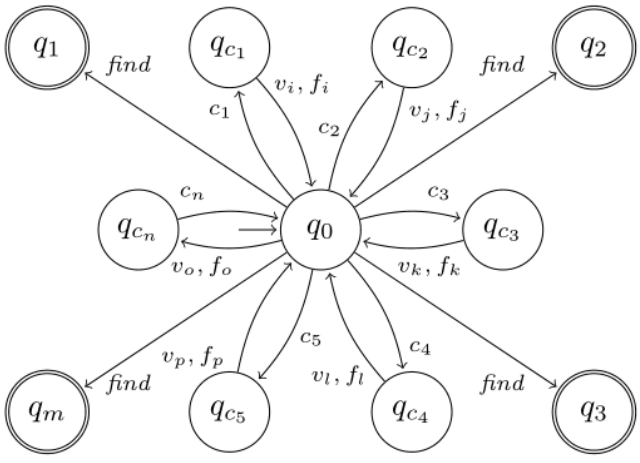

Figura 1. Autômato adaptativo *M* do modelo.

O autômato adaptativo *M* recebe cadeias de entrada que representam as consultas sobre o conjunto de dados. O tipo de consulta a ser submetida ao modelo parte dos objetivos da extração e da definição do problema.

Considere os conjuntos $C = \{c_1, c_2, \ldots, c_n\}$ de características e $V = \{v_1, v_2, \ldots, v_i\}$ de valores associados às características. O autômato adaptativo *M* da Figura 1 possui o estado inicial q_0, *n* estados intermediários, onde $n \in \mathbb{N}$ é o total de características existentes no conjunto de dados, e *m* estados finais, $m \in \mathbb{N}$ é o total de indivíduos disponíveis. Os estados intermediários q_i são alcançáveis a partir de q_0 através de uma transição consumindo a característica $c_1 \in C$, além de haver uma transição de volta consumindo um determinado valor $v_j \in V$ associado à característica c_i e executando uma determinada função adaptativa f_k correspondente. Finalmente, o autômato possui *m* transições consumindo o símbolo especial <*find*> partindo de q_0 até os estados finais. O alfabeto utilizado é definido como $\Sigma = C \cup V \cup \{<find>\}$. A cadeia *w* submetida ao autômato deve seguir o formato $w = (cv) * <find>$, $c \in C$, $v \in V$, ou seja, procuram-se indivíduos que contêm os pares <*característica*>=<*valor*> seguido do símbolo especial <*find*>, que efetivamente realiza a busca no autômato.

O autômato adaptativo *M* do modelo é definido de forma não-determinística, a princípio. Semanticamente, o indeterminismo denota os indivíduos candidatos à escolha. No início do processo de reconhecimento da cadeia *w*, todos os indivíduos têm a mesma probabilidade de escolha, portanto todos os estados finais que os representam são alcançáveis a partir do estado inicial q_0 consumindo o símbolo especial <*find*>.

Do ponto de vista de extração de conhecimento a partir do conjunto de dados disponível, o quão indeterminístico o autômato adaptativo for após a submissão e reconhecimento da cadeia de entrada tem relação direta com a entropia do conjunto de indivíduos disponível. Ao longo do processo de reconhecimento da cadeia, espera-se que o indeterminismo seja gradativamente reduzido, através da filtragem dos indivíduos candidatos à escolha. Através do modelo adaptativo, a busca por características e a eventual identificação de indivíduos ocorre em tempo real. O modelo é compacto e simplificado, o que favorece a visualização e interpretação dos resultados obtidos.

A topologia do autômato adaptativo *M* ao final do reconhecimento da cadeia de entrada indica semanticamente qual foi o conhecimento extraído. Se a cadeia não foi aceita – não há estados finais alcançáveis, independente da situação de indeterminismo ou não do autômato – não foi possível identificar indivíduos de acordo com os parâmetros definidos na cadeia. Caso *M* seja determinístico e a cadeia foi aceita, o estado final $q_i \in F$ alcançável indica que o indivíduo *i* foi univocamente identificado. Quando *M* é indeterminístico e a cadeia foi aceita, todos os estados $\{q_i, q_j, \ldots, q_k\} \subseteq F$ alcançáveis representam um grupo de indivíduos com as características procuradas.

O indeterminismo permite a visualização e definição de grupos de interesse na extração de conhecimento de modo incremental. Não é necessário filtragem e pós-processamento dos resultados, pois a topologia resultante do autômato garante a consistência relacional dos dados obtidos. As funções adaptativas do modelo realizam "cortes" no autômato, removendo estados e transições que não aderem aos objetivos definidos na cadeia de entrada para extração de conhecimento. A adaptatividade permite a redução e simplificação do conjunto de dados até o nível desejado.

Como exemplo, considere o conjunto de dados da Tabela 1. Cada um dos cinco indivíduos tem suas quatro características mapeadas. O autômato adaptativo correspondente é ilustrado na Figura 2. Observe que a coluna *nome*, por tratar-se da identidade de cada indivíduo, é associada aos estados finais.

Tabela I. Exemplo de conjunto de dados para ilustração do modelo adaptativo.

id	nome	sexo	olhos	cabelo	estatura
1	Ana	feminino	azuis	castanho	alta
2	José	masculino	azuis	preto	alta
3	João	masculino	castanhos	loiro	mediana
4	Maria	feminino	verdes	loiro	baixa
5	Pedro	masculino	castanhos	ruivo	baixa

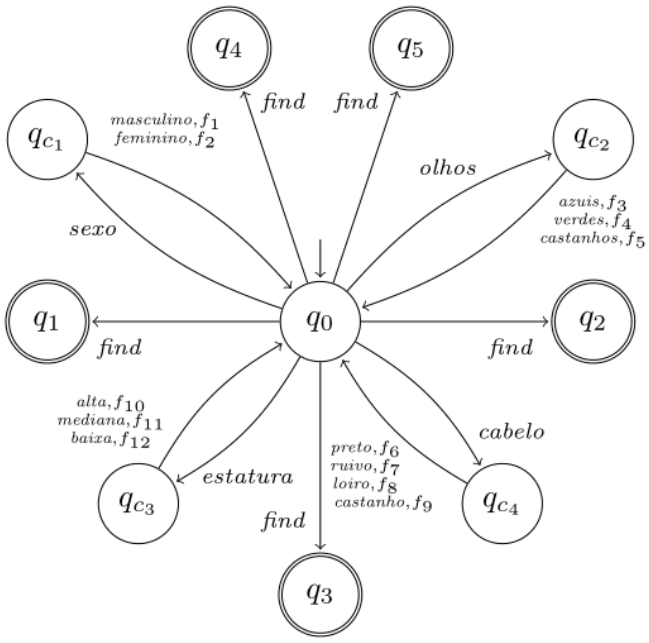

Figura 2. Autômato *M* representando os dados da Tabela I.

Deseja-se, por exemplo, identificar todos os indivíduos que possuem olhos verdes. Para tal, a cadeia de entrada $w_1 = <olhos><verdes><find>$ é submetida ao autômato adaptativo *M*. Como w_1 foi aceita por *M*, o conjunto de dados contém indivíduos que possuem as características informadas. O autômato resultante é ilustrado na Figura Figura 3.

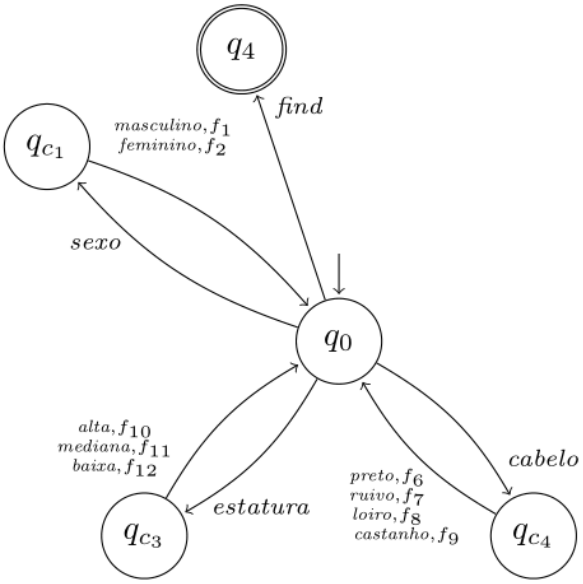

Figura 3. Autômato adaptativo *M* resultante após a submissão da cadeia $w_1 =$ <olhos><verdes><find>.

De acordo com a Figura 3, o autômato adaptativo *M* resultante é determinístico, significando que um único indivíduo foi encontrado. O estado final alcançável é q_4, associado a *Maria*, o único indivíduo da Tabela I que possui olhos verdes.

Em outro exemplo, deseja-se identificar todos os indivíduos que possuem cabelo loiro. Para tal, a cadeia de entrada $w_2 =$ <cabelo><loiro><find> é submetida ao autômato adaptativo *M*. O autômato resultante é ilustrado na Figura 4.

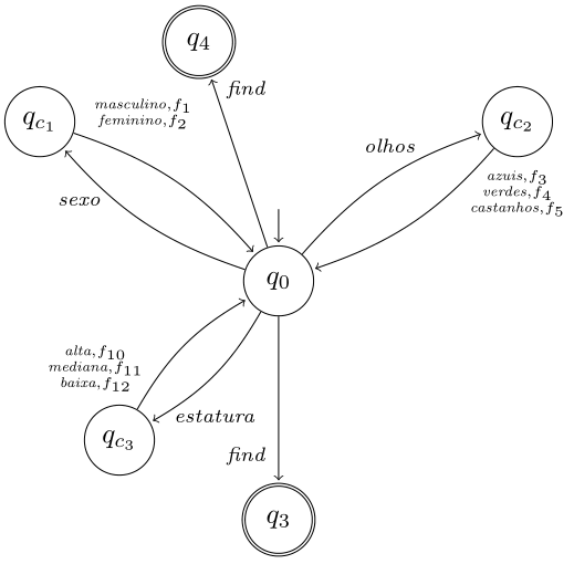

Figura 4. Autômato adaptativo *M* resultante após a submissão da cadeia $w_2 =$ <cabelo><loiro><find>.

De acordo com a Figura 4, o autômato adaptativo *M* resultante é indeterminístico, significando que um grupo de indivíduos com as mesmas características foi encontrado. Os estados finais alcançáveis são q_3 e q_4, associados a *João* e *Maria*. Caso a intenção seja obter apenas as mulheres com cabelo loiro, acrescenta-se um par <*característica*>=<*valor*> à cadeia, de modo que $w_3 =$ <cabelo><loiro><sexo><feminino><find>. Neste caso, somente o indivíduo *Maria* é encontrado.

A mineração adaptativa de dados pode favorecer o processo de descoberta de conhecimento em bancos de dados através de modelos adaptativos simplificados e consistentes. Como se pode intuir a partir da ilustração apresentada, a utilização da área de tecnologia adaptativa permite a extração incremental de conhecimento, de forma eficiente.

IV. Experimento e Análise

A mineração adaptativa de dados apresenta-se como uma alternativa atraente à mineração de dados tradicional. É importante destacar que mineração adaptativa e a tradicional não são mutuamente exclusivas, ao contrário, são completamente compatíveis uma com a outra, podendo por isso ser usadas simultaneamente, de acordo com a conveniência em cada caso. Esta seção descreve um experimento proposto para coletar dados estatísticos acerca do desempenho do modelo adaptativo e da consistência desses dados.

A primeira parte do experimento consistiu na definição e implementação de um simulador para a identificação de indivíduos através do modelo adaptativo ilustrado anteriormente. O simulador possui as seguintes funcionalidades:

a) geração, leitura e mapeamento de dados, em memória ou disco rígido, para automatizar o gerenciamento e processamento do conhecimento a ser extraído,
b) definição do modelo de autômato adaptativo correspondente ao problema, de acordo com o conjunto de dados disponível ao simulador,
c) terminal de consulta na forma de cadeias de entrada de tamanho arbitrário submetidas ao autômato adaptativo, no formato requerido apresentado na Seção III,
d) geração de *snapshots* da configuração do autômato para cada símbolo consumido, representados através de arquivos no formatos `dot`[1] e `png`[2],
e) geração de gráficos de análise dos dados, baseado no reconhecimento das cadeias de entrada, no tempo de execução e no cálculo dos bits de informação.

O simulador foi escrito utilizando-se a linguagem Python e executado através de linha de comando em um ambiente Linux de 64 bits.

As amostras utilizadas neste experimento foram geradas através do simulador. Inicialmente, foram gerados milhares de conjuntos de dados aleatórios, de tamanhos arbitrários, com registros no formato apresentado na Tabela II. Todas as amostras geradas possuem características estatísticas controladas e bem definidas.

[1] Linguagem de descrição de grafos em texto puro.
[2] Formato de dados utilizado para imagens.

Tabela II. Formato de registro para geração dos conjuntos de dados.

id	sexo	olhos	cabelos	estatura	pele	estado civil	idade
-	masc	verdes	cast.	baixa	clara	solt.	criança
-	fem	azuis	preto	med.	morena	cas.	jovem
-	-	cast.	ruivo	alta	negra	-	adulto
-	-	-	loiro	-	-	-	idoso

A partir dos conjuntos gerados, foram escolhidos três deles, com os seguintes tamanhos: $n(C_1) = 10$, $n(C_2) = 100$ e $n(C_3) = 1000$ indivíduos. Cada conjunto escolhido representou uma determinada população a ser analisada. Para cada indivíduo, foram geradas sete características e um identificador único para verificação e validação dos dados, de acordo com a Tabela II.

Na segunda parte do experimento, após a definição dos três conjuntos de dados, foram realizadas medições de tempo de execução e consistência dos dados durante as consultas aos autômatos, de acordo com as tarefas definidas na Tabela III.

Tabela III. Tarefas principais do experimento.

ID	Descrição
t_1	Tempo de execução das consultas sob os conjuntos C_1, C_2 e C_3
t_2	Consultas para identificação única de indivíduos
t_3	Consultas para identificação de grupos de indivíduos

De modo complementar, outras medições relacionadas a alguns métodos tradicionais de extração de conhecimento foram realizadas, de acordo com as tarefas definidas na Tabela IV.

Tabela IV. Tarefas complementares do experimento.

ID	Descrição
t_4	Produto cartesiano das características, calculada a taxa dos bits de informação comparativamente ao valor limite dos bits de informação do conjunto de dados
t_5	Maior valor para cada combinação das características, de modo comparativo ao valor limite dos bits de informação
t_6	Cálculo dos bits de informação dos conjuntos C_1, C_2 e C_3

A Figura 5 ilustra o tempo de execução das consultas sob os conjuntos C_1, C_2 e C_3. O simulador gerou mil consultas de tamanho arbitrário, variando entre um e quinze símbolos, que foram submetidas na forma de cadeias de entrada, de acordo com o formato da cadeia de entrada apresentado na Seção III, aos três autômatos correspondentes. O eixos *x* e *y* representam o número de pares <*característica*>=<*valor*> consultados na cadeia e o tempo de execução em segundos, respectivamente.

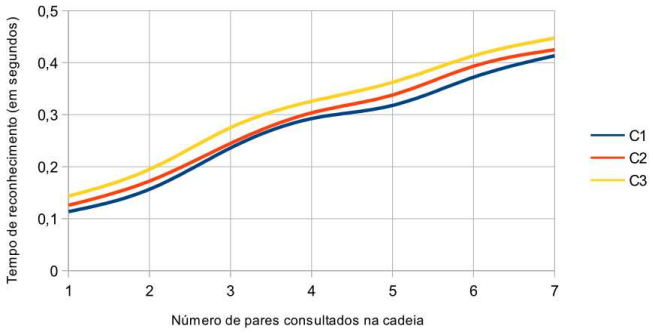

Figura 5. Tempo de execução das consultas sob os conjuntos C_1, C_2 e C_3.

De acordo com a Figura 5, o tempo de reconhecimento das cadeias de entrada para os três conjuntos analisados mostrou-se linear, proporcional ao comprimento da cadeia. É possível observar que não houve uma variação significativa entre os tempos dos três conjuntos, apesar da diferença entre o total de elementos de cada conjunto.

Utilizando as consultas geradas no teste anterior, foram verificados ainda os tempos de execução de consultas que resultavam em identificação única – apenas um estado final alcançável, gerando portanto um autômato determinístico – e grupo de indivíduos – dois ou mais estados finais alcançáveis, preservando o indeterminismo do autômato. Os resultados são ilustrados na Figura 6. É importante destacar que, para este teste, os resultados foram normalizados para que não houvesse discrepância entre os tamanhos das consultas e os tempos de execução.

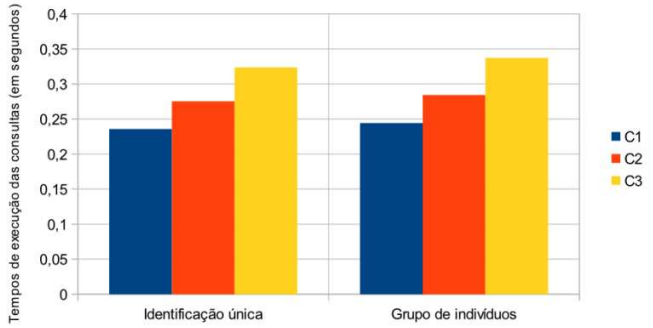

Figura 6. Tempos de execução das consultas para identificação única e de grupo.

De acordo com a Figura 6, é possível observar que os autômatos apresentaram um tempo semelhante para o reconhecimento das cadeias de entrada, não fazendo distinção do tipo de identificação retornada – única ou grupo. Nota-se que o modelo proposto aparentemente não gera sobrecarga de recursos computacionais ao retornar um grupo arbitrário de indivíduos – o tempo é praticamente o mesmo, independente do resultado retornado.

A Figura 7 ilustra o produto cartesiano das características presentes em cada um dos conjuntos e seus respectivos bits de informação. Alguns dos valores nulos foram omitidos para facilitar a visualização, uma vez que não interferem na interpretação dos resultados.

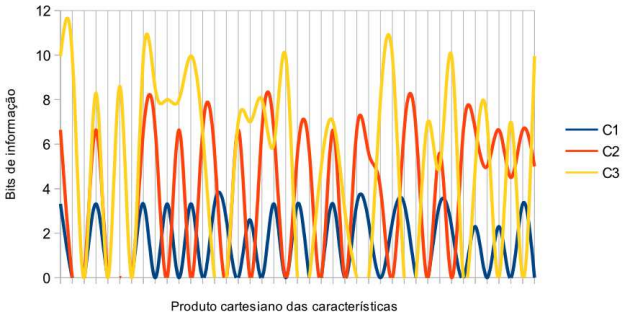

Figura 7. Produto cartesiano das características presentes em cada conjunto e seus respectivos bits de informação.

De acordo com a Figura 7, os conjuntos C_1 e C_2 apresentam mais picos do que o conjunto C_3. Dado o tamanho do conjunto C_3 e os registros no formato definido na Tabela II, este possui maior grau de entropia do que os outros conjuntos, portanto torna-se mais difícil a identificação única de um indivíduo. O grau de entropia aumenta ou diminui de acordo com o tamanho da amostra. Pode-se inclusive falar em *taxa de redução de indeterminismo*, inversamente proporcional ao grau de entropia do conjunto. É importante observar que o valor máximo dos bits de informação é diferente para cada conjunto.

A Figura 8 ilustra os maiores valores de bits de informação encontrados para cada combinação de pares <*característica*>=<*valor*> de cada conjunto, comparados ao valor limite dos bits de informação.

Figura 8. Maiores valores para cada combinação das características de cada conjunto comparados ao valor limite dos bits de informação.

De acordo com a Figura 8, é possível verificar a influência do tamanho da amostra para a identificação única de indivíduos. O conjunto C_1, por possuir apenas 10 indivíduos, permite a identificação única de um indivíduo com apenas um par <*característica*>=<*valor*>; o conjunto C_3 contendo 1000 indivíduos, no entanto, requer um refinamento maior na cadeia de entrada para a identificação única. No modelo adaptativo proposto, a relação de entropia e indeterminismo está associada ao tamanho do volume de dados disponível e a quantidade de características.

Como teste adicional, o simulador gerou um conjunto especial contendo 1000 indivíduos com 32 características cada. Durante a reprodução dos testes anteriores, o autômato adaptativo do modelo proposto ainda apresentou sua média de execução abaixo de 0,5 segundos, mesmo para cadeias de entrada com comprimento expressivo – por exemplo, 65 símbolos. Não foi possível reproduzir o mesmo teste adicional utilizando uma técnica tradicional de extração de conhecimento devido à explosão combinatória do algoritmo, gerando aproximadamente $1{,}83 \times 10^{19}$ combinações, exaurindo os recursos computacionais, demandando tempo impraticável e inviabilizando o seu uso. Outras técnicas tradicionais não foram aplicadas por serem genéricas demais para o problema em questão.

V. Discussões

Esta seção trata das discussões acerca do modelo adaptativo para identificação de indivíduos e dos conceitos de mineração adaptativa de dados. A seguir, são apresentados alguns questionamentos.

Os resultados obtidos sobre o tempo de execução do modelo adaptativo para extração de conhecimento, proposto neste artigo, apresentam-se favoráveis, dado que uma parcela significativa das técnicas tradicionais de mineração de dados apresenta tempo exponencial para extração e análise dos dados, o que onera o processo de mineração e inviabiliza sua utilização para volumes de dados com tamanho considerável. Entretanto, é importante notar que seriam necessárias mais amostras de conjuntos de dados de tamanho arbitrário e um comparativo entre as técnicas de mineração de dados existentes para atestar de modo definitivo a eficiência do modelo adaptativo.

Foi possível verificar, através do experimento descrito na Seção IV, que o reconhecimento em tempo quase linear da cadeia de entrada submetida ao autômato adaptativo do modelo permite uma visualização incremental contínua do processo de extração. Além disso, dado o rigor da definição formal do autômato adaptativo, os dados e a topologia que os representa apresentam uma consistência contínua, mantida durante todo o processo até o consumo do último símbolo da cadeia. A consistência de tais dados permite uma homogeneidade do modelo. Na mineração de dados tradicional, é necessário aguardar o término das fases descritas na Seção III para efetivamente analisar quaisquer dados obtidos e extrair conhecimento consistente, eventualmente através de um especialista de domínio.

Embora os testes de tempo de execução demonstrem uma eficiência considerável do modelo adaptativo, é importante destacar que este foi definido para solucionar um problema específico – a identificação de indivíduos. A mineração adaptativa de dados pode requerer um modelo para cada problema, ao contrário da mineração de dados tradicional, na qual seus modelos e técnicas são mais abrangentes. Além disso, a modelagem adaptativa pode tornar-se complexa comparativamente aos parâmetros e variáveis das técnicas tradicionais.

O modelo adaptativo para identificação de indivíduos é genérico o suficiente para ser utilizado em diversas aplicações, tais como motor de estratégias para jogos de tabuleiro, seleção de candidatos, busca de currículos, formação de equipes esportivas, otimização de problemas de logística, escolha de estratégias comerciais, *marketing* direcionado, análise de perfil populacional, otimização de campanhas de vacinação,

distribuição de recursos, identificação de criminosos, mapas territoriais, identificação de plágios em trabalhos acadêmicos e análises ambientais. Além das aplicações diretas, o modelo adaptativo oferece subsídios para a definição de um banco de dados adaptativo, otimizado para grandes consultas e para a crescente área de computação em nuvem.

VI. Conclusões

Este artigo apresentou uma proposta de modelo adaptativo para identificação de indivíduos, juntamente com o conceito de mineração adaptativa de dados. O modelo adaptativo mostrou-se viável para a utilização em grandes volumes de dados, permitindo a identificação única de indivíduos ou retornando grupos de interesse, sendo sua aplicação extensível para as mais diversas áreas. O tipo de tratamento utilizado para a identificação de indivíduos pode servir como elo de aproximação entre métodos estatísticos e métodos discretos, e que o seu estudo poderá trazer luzes sobre a utilização de métodos adaptativos – que são discretos – no estudo de fenômenos contínuos. Isso também ajudaria a facilitar o uso de tratamentos discretos para casos usualmente tratados estatisticamente, como acontece em algumas situações no caso do processamento de linguagem natural.

O conceito de mineração adaptativa de dados, introduzido neste artigo, pode tornar-se uma alternativa viável aos métodos existentes para extração de conhecimento. A crescente área de computação em nuvem pode beneficiar-se de técnicas e modelos adaptativos para oferecer soluções de baixo custo, alto poder computacional e suporte a grandes volumes de dados. Além disso, a idéia de um banco de dados adaptativo é desafiadora, podendo gerar contribuições para áreas correlatas, como otimização, recuperação da informação e mineração de dados.

A área de tecnologia adaptativa pode proporcionar soluções consistentes que atendam às necessidades inerentes a cada aplicação. A utilização de dispositivos adaptativos para a definição de técnicas de extração de conhecimento pode beneficiar não somente o processo de extração em si, mas permitir a obtenção de outros dados que não estão disponíveis através dos métodos tradicionais.

Referências

[1] Jess3, "The social universe," Jess3 Data Visualization Agency, Tech. Rep., 2010.

[2] H. Blockeel and M. Sebag, "Scalability and efficiency in multi-relational data mining," *SIGKDD Explorations*, vol. 5, pp. 17–30, 2003.

[3] M. A. A. Sousa and A. H. Hirakawa, "Robotic mapping and navigation in unknown environments using adaptive automata," in *Proceedings of International Conference on Adaptive and Natural Computing Algorithms – ICANNGA 2005*, Coimbra, Portugal, March 21–23 2005.

[4] H. Pistori and J. J. Neto, "Adaptree – proposta de um algoritmo para indução de Árvores de decisão baseado em técnicas adaptativas," in *Anais da Conferência Latinoamericana de Informática – CLEI 2002*, Montevideo, Uruguai, Novembro 2002.

[5] T. Pedrazzi, A. H. Tchemra, and R. L. A. Rocha, "Adaptive decision tables – a case study of their application to decision-taking problems," in *Proceedings of International Conference on Adaptive and Natural Computer Algorithms – ICANNGA 2005*, 2005.

[6] C. Menezes and J. J. Neto, "Um método para a construção de analisadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos," in *V PROPOR – Encontro para o processamento computacional de Português falado e escrito*, 2000.

[7] J. Luz and J. J. Neto, "Tecnologia adaptativa aplicada à otimização de código em compiladores," in *IX Congreso Argentino de Ciencias de La Computación – CACIC 2003*, 2003.

[8] P. R. M. Cereda, "Modelo de controle de acesso adaptativo," Dissertação de Mestrado, Departamento de Computação, Universidade Federal de São Carlos, 2008.

[9] ____, "Servidor web adaptativo," in *WTA 2010: Workshop de Tecnologia Adaptativa*, Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil, 2010.

[10] R. L. A. Rocha and J. J. Neto, "Uma proposta de linguagem de programação funcional com características adaptativas," in *IX Congreso Argentino de Ciencias de la Computación*, 2003.

[11] A. V. Freitas and J. J. Neto, "Adaptive languages and a new programming style," in *6th WSEAS International Conference on Applied Computer Science*, 2006.

[12] A. R. Camolesi and J. J. Neto, "Modelagem adaptativa de aplicações complexas," in *Conferência Latinoamericana de Informática – CLEI*, 2004.

[13] A. R. Camolesi, "Proposta de um gerador de ambientes para modelagem de aplicações usando tecnologia adaptativa," Ph.D. dissertation, Escola Politécnica da USP, 2007.

[14] H. Pistori, J. J. Neto, and M. C. Pereira, "Adaptive non-deterministic decision trees: general formulation and case study," *INFOCOMP – Journal of Computer Science*, 2006.

[15] C. Bravo, J. J. Neto, F. S. Santana, and A. M. Saraiva, "Towards an adaptive implementation of genetic algorithms," in *Conferência Latinoamericana de Informática – CLEI*, 2007.

[16] J. J. Neto, "Um levantamento da evolução da adaptatividade e da tecnologia adaptativa," *IEEE Latin America Transactions*, vol. 5, no. 7, pp. 496–505, 2007.

[17] W. Aiello and P. McDaniel, *Lecture 1, Intro: Privacy*, Stern School of Business, NYU, 2004.

[18] Electronic Frontier Foundation, "Annual report 2009–2010," Electronic Frontier Foundation, Tech. Rep., 2010.

[19] Brasil, *Código Penal*, Brasília, Distrito Federal, 1940.

[20] M. Teltzrow and A. Kobsa, "Communication of privacy and personalization in e-business," in *Workshop Wholes: A multiple view of individual privacy in a networked world*, Stockholm, Sweden, 2004.

[21] S. Azambuja, "Estudo e implementação da análise de agrupamento em ambientes virtuais de aprendizagem," Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, 2005.

[22] M. Koch and K. Moeslein, "User representation in ecommerce and collaboration applications," in *BLED 2003 Proceedings*, 2003.

[23] D. Kristol and L. Montulli, "HTTP state management mechanism," Bell Laboratories, Lucent Technologies, RFC 2965, 2000.

[24] A. L. Montgomery, "Using clickstream data to predict www usage," University of Maryland, Tech. Rep., 2003.

[25] S. Sae-Tang and V. Esichaikul, "Web personalization techniques for e-commerce," in *Active Media Technology: 6th International Computer Science Conference*, 2005.

[26] D. L. Lee, M. Zhu, and H. Hu, "When location-based services meet databases," *Mob. Inf. Syst.*, vol. 1, pp. 81–90, 2005.

[27] J. Raper, G. Gartner, H. Karimi, and C. Rizos, "Applications of location-based services: a selected review," *J. Locat. Based Serv.*, vol. 1, pp. 89–111, 2007.

[28] J. Paay and J. Kjeldskov, "Understanding the user experience of location-based services: five principles of perceptual organisation applied," *J. Locat. Based Serv.*, vol. 2, pp. 267–286, 2008.

[29] K. Wac and L. Ragia, "Lspenv: location-based service provider for environmental data," *J. Locat. Based Serv.*, vol. 2, pp. 287–302, 2008.

[30] J. Raper, G. Gartner, H. Karimi, and C. Rizos, "A critical evaluation of location based services and their potential," *J. Locat. Based Serv.*, vol. 1, pp. 5–45, 2007.

[31] G. Ghinita, "Private queries and trajectory anonymization: a dual perspective on location privacy," *Trans. Data Privacy*, vol. 2, pp. 3–19, 2009.

[32] J. Lukás, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, pp. 205–214, 2006.

[33] E. Y. L. Kai San Choi and K. K. Wong, "Source camera identification using footprints from lens aberration," *SPIE-IS&T Electronic Imaging*, vol. 6069, 2006.

[34] O. Hilton, "The complexities of identifying the modern typewriter," *Journal of Foresinc Sciences*, vol. 2, 1972.

[35] T. Kohno, A. Broido, and K. Claffy, "Remote physical device fingerprinting," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, 2005.
[36] IETF, "Hypertext Transfer Protocol – HTTP/1.0," Internet Engineering Task Force, Tech. Rep., 1996.
[37] P. Eckersley, "How unique is your web browser?" Electronic Frontier Foundation, Tech. Rep., 2009.
[38] M. Ackerman and L. F. Cranor, "Privacy critics – safeguarding users' personal data," 1999.
[39] D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Commun. ACM*, vol. 24, no. 2, pp. 84–90, 1981.
[40] I. Goldberg, D. Wagner, and E. Brewer, "Privacy-enhancing technologies for the internet," in *IEEE Spring Computer Conference*. San Jose, CA, USA: IEEE Computer Society Press, February 1997, pp. 103–110.
[41] D. Goldschlag, M. Reed, and P. Syverson, *Anonymous Connections and Onion Routing*, Assurance Computer Systems, Naval Research Laboratory, Washington, DC, 1996.
[42] U.S. Census Bureau, "Total population of the world," U.S. Census Bureau, Tech. Rep., 2011.
[43] P. Eckersley, "A primer on information theory and privacy," Electronic Frontier Foundation, Tech. Rep., 2010.
[44] H. Jiawei and M. Kamber, *Data Mining: Concept and Techniques*. Morgan Kaufmann, 2006.
[45] H. Mannila, "Data mining: machine learning, statistics and databases," *International Conference on Statistics and Scientific Database Management*, 1996.
[46] U. Fayyad, G. Piatetsky-shapiro, P. Smyth, and T. Widener, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, pp. 27–34, 1996.
[47] J. J. Neto, "Contribuições à metodologia de construção de compiladores," Tese de Livre Docência, Escola Politécnica da Universidade de São Paulo, São Paulo, 1993.
[48] ____, "Adaptive automata for context-dependent languages," *SIGPLAN Notices*, vol. 29, no. 9, pp. 115–124, 1994.

Paulo Roberto Massa Cereda é graduado em Ciência da Computação pelo Centro Universitário Central Paulista (2005) e mestre em Ciência da Computação pela Universidade Federal de São Carlos (2008). Atualmente, é desenvolvedor de software, atuando em diversos segmentos de mercado, com enfoque principal em software livre. É membro ativo do repositório de código fonte *SourceForge* desde 2005, contribuindo com bibliotecas e softwares de propósito geral. Tem experiência na área de Ciência da Computação, atuando principalmente nos seguintes temas: inteligência artificial, linguagens de programação, teoria da computação e tecnologia adaptativa.

João José Neto é graduado em Engenharia de Eletricidade (1971), mestre em Engenharia Elétrica (1975), doutor em Engenharia Elétrica (1980) e livre-docente (1993) pela Escola Politécnica da Universidade de São Paulo. Atualmente, é professor associado da Escola Politécnica da Universidade de São Paulo e coordena o LTA – Laboratório de Linguagens e Tecnologia Adaptativa do PCS – Departamento de Engenharia de Computação e Sistemas Digitais da EPUSP. Tem experiência na área de Ciência da Computação, com ênfase nos Fundamentos da Engenharia da Computação, atuando principalmente nos seguintes temas: dispositivos adaptativos, tecnologia adaptativa, autômatos adaptativos, e em suas aplicações à Engenharia de Computação, particularmente em sistemas de tomada de decisão adaptativa, análise e processamento de linguagens naturais, construção de compiladores, robótica, ensino assistido por computador, modelagem de sistemas inteligentes, processos de aprendizagem automática e inferências baseadas em tecnologia adaptativa.