Emprego de adaptatividade em mineração de dados jurídicos - Uma primeira análise

Leme, Pedro C.
Departamento de Engenharia de Sistemas Eletrônicos
Escola Politécnica da USP
São Paulo, Brasil
pedro.leme@yahoo.com

***Resumo*—A extração de informações de dados judiciais públicos no Brasil, surgida principalmente a partir de 2006, tem se tornado uma área de crescente visibilidade. O uso de robôs de coleta desses dados a partir de diferentes sistemas e sites requer codificação específica e orquestração de funcionamento baseado em diversos fatores, travas e características de cada site. Devido a essas características o controle de muitos robôs e máquinas de maneira quase autônoma se faz necessário, não sendo mais possível um ajuste manual para cada tipo de sistema-alvo de maneira satisfatória. Pretende-se analisar o cenário e futuramente aplicar os conceitos de Adaptatividade no desenvolvimento de um mecanismo adaptativo que irá atuar sobre o subsistema de controle de instâncias de modo que o funcionamento deste passe a requerer menor envolvimento humano e apresente melhora em qualidade.**

Palavras-chave—adaptatividade; mineração de dados; aplicações

I. Introdução

Com o alto nível de judicialidade[1] e a determinação do Conselho Nacional de Justiça (CNJ)[2] de digitalizar os processos judiciais no país todo a partir de 2006, surgiu uma área de atuação, já explorada em países como os EUA, especializada na extração, compilação e geração de valor de dados legais e judiciais públicos no Brasil.

A empresa Digesto[3] surgiu em 2012 a partir dessa ideia, e hoje extrai dados diariamente de 500 fontes de diários oficiais e mais de 50 tribunais em todo o território brasileiro de maneira digital e automática, tendo 100 milhões de processos e mais de dois bilhões de andamentos processuais, dentre outras bases. A extração dos dados se dá a partir de sites e portais de tribunais e órgãos públicos brasileiros, cujos dados são abertos por lei.

Robôs e sistemas desenvolvidos especialmente para cada site ou sistema realizam as requisições via método HTTP/S ou *webservices*, depois fazem o *parsing* do retorno e análise dos dados retornados, quer seja HTML, JSON, XML ou texto puro, para a extração de informações que são então tratadas e enviadas para um banco de dados relacional. Essa informação depois é usada para geração de relatórios e inferência de classificações diversas via aprendizado por máquina a fim de obter e inserir inteligência sobre esses dados.

II. Objetivo

Cada sistema usado pelos diferentes órgãos públicos têm características diferentes, seja em termos de caminho de acesso, estrutura interna, apresentação dos dados, horários de funcionamento, travas e controles de número de acesso, necessidade ou não de *login* além de instabilidade do serviço, e isso se reflete nas características dos robôs (*workers*), sendo necessárias diretivas específicas para cada um, e tais parâmetros são atualmente codificados num sistema de execução e ajustados manualmente a partir de observações e controles manuais, diários ou semanais.

Esse sistema, responsável por orquestrar a extração dos dados, tem a capacidade de iniciar e paralisar as máquinas para cada robô, decidir o horário em que devem trabalhar, além de configurar parâmetros como número de processos e *threads* simultâneas e estratégia de coleta de dados e controle de filas de pedidos de informação com base nas variáveis inseridas nele por um operador.

Com o crescimento dos serviços, robôs e volume de dados obtidos diariamente, o controle manual desses parâmetros tem se mostrado cada vez mais complexo e demorado, além de apresentar resultados muitas vezes aquém do esperado, resultando em falta de dados por tempo inadequado de execução de determinado robô, gasto excessivo com máquinas ligadas esperando por tarefas ou ainda problemas mais urgentes, como bloqueio de IP's e *logins* em sites devido a grandes volumes de requisições em um curto espaço de tempo.

O objetivo do estudo é fazer um levantamento do cenário atual e analisar as possibilidades de uso da Adaptatividade no sistema de gerenciamento dos robôs de extração, de modo que as características dos sistemas e análise do funcionamento destes em tempo real possibilitem a troca automática das estratégias usadas no acesso a cada um dos sites e portais, melhorando a qualidade dos dados extraídos ao mesmo tempo em que aumenta sua quantidade e reduz custos com infraestrutura.

III. Infraestrutura e fluxo atual de dados

O fluxo do sistema inicia-se com robôs que diariamente coletam e extraem informações de diários oficiais do país todo. Esses robôs são ligados automaticamente por agendamento e através de requisições HTTP e *parsing* de HTML baixam arquivos, em sua maioria pdf, dos diários oficiais de tribunais

de todo o Brasil. Nos diários constam pequenos trechos de andamentos e solicitações de processos, assim como uma relação de todos os processos novos criados no dia anterior.

Com tais dados em mãos um sistema extrai dos textos números num formato específico, compatíveis com o padrão de numeração única instituído pelo CNJ, e os envia como mensagens para um sistema de filas específicas para cada tribunal, o mesmo de onde cada número foi extraído do respectivo diário.

A partir do enfileiramento de mensagens com os números num sistema de mensageria, outros robôs, especializados em extração de dados a partir de páginas web, começam a consumir esses números e a fazer as requisições nos tribunais de justiça, a fim de obter informações mais detalhadas acerca dos processos e seus andamentos. Tal extração faz uso de requisições HTTP/S e de *webservices*, além de *parsing* de HTML e geração e navegação em árvores estruturadas a partir deste. Encontrando e absorvendo essas informações dos dados devolvidos pelo tribunal, os robôs realizam uma limpeza e normalização para depois inseri-los em um banco de dados relacional, capaz de armazená-los e garantir suas propriedades e conexões.

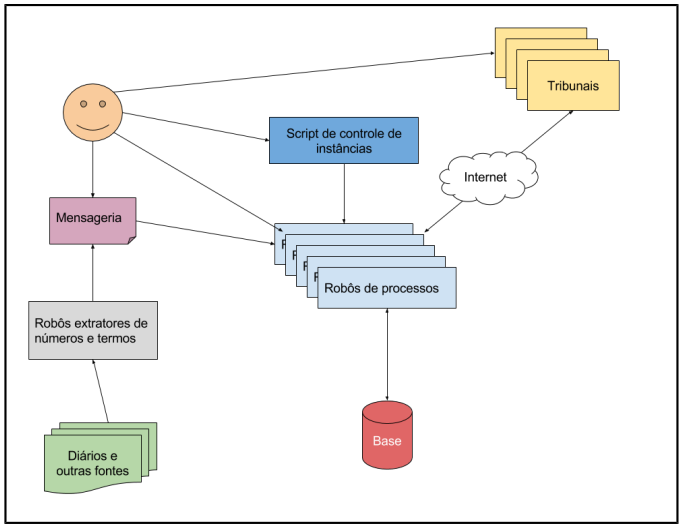

Fig. 1. Representação simplificada do sistema atual

Da representação simplificada do sistema atual, na Figura 1, nota-se que os robôs têm seu funcionamento orquestrado por um *script* de controle de instâncias que armazena parâmetros de configuração, tempo de execução e quantidade de instâncias de cada *worker*, para então ligá-las e desligá-las automaticamente seguindo essas variáveis. Atualmente os parâmetros são configurados manualmente por um operador, que monitora o funcionamento dos robôs, a quantidade de mensagens nas filas e o andamento das requisições nos tribunais a fim de decidir a quantidade de instâncias de cada robô e seus horários de execução para ajuste individual no curto prazo.

IV. Dificuldades do cenário

Com um número pequeno de sites de tribunais e robôs o controle manual das variáveis é relativamente rápido e efetivo, pois é possível acompanhar em tempo real o funcionamento e ajustar os parâmetros praticamente *on-the-fly* no *script* de controle. No entanto, com o crescimento da quantidade e cobertura dos *workers*, tal procedimento tem se tornado custoso e sujeito a falhas, quer seja pela alta frequência de modificações nos robôs ou pela grande variação no funcionamento de cada um deles.

Um ponto a se considerar no desenvolvimento de robôs específicos para uma grande gama de sites e serviços diferentes é a quantidade de características existentes, assim como o número expressivo de possibilidades de coexistência de uma ou mais destas características num mesmo robô. Tem-se como exemplos a existência de *captcha*, necessidade de *login*, horários específicos de acesso, uso de diferentes rotas dependendo do dado que se busca, controle de número de requisições concorrentes ou diárias, bloqueio de acesso a determinadas informações e principalmente instabilidade e indisponibilidade de diversos sites e serviços públicos.

Nos robôs em funcionamento atualmente boa parte dessas características já é levada em conta, quer seja pelos mecanismos de acesso com *login* e detecção de falhas e indisponibilidade, mas pelos sites-alvo serem sistemas sobre os quais não há qualquer conhecimento interno do funcionamento ou decisão dos mantenedores, os robôs e o sistema de controle não é preparado para tomar qualquer medida minimamente inteligente de maneira automática caso alguma dessas características mude.

Há de se juntar a esse problema de modificação sem aviso prévio o fato de haver flutuação nas quantidades de mensagens de pedidos nas filas, quer seja por algum fator externo, como um dia sem publicação de diário oficial ou publicação dupla ou um novo pedido com números já conhecidos que precisem de atualização, ou mesmo interno, como um ajuste de horário e quantidade de máquinas errado que aos poucos deixe acumular uma grande quantidade de mensagens a serem consumidas até que fuja do controle.

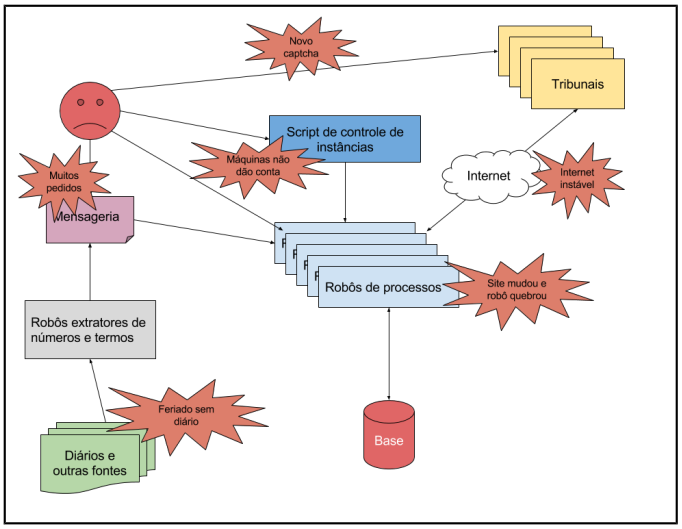

Fig. 2. Ilustração de problemas frequentes no funcionamento dos sistemas

Com a adição de todos os problemas possíveis e conhecidos que ocorrem numa escala semanal, senão diária, multiplicados

pela quantidade de sistemas-alvo na cobertura de extração, o trabalho manual de ajuste do *script* de controle de instâncias já não apresenta eficiência aceitável, e um novo mecanismo para esse subsistema na infraestrutura é necessário.

V. Adaptatividade

O conceito de Adaptatividade, conforme especificado por Neto, J. J.[4], descreve um sistema no qual um dispositivo adjacente é capaz de ter seu conjunto de regras modificado por um mecanismo adaptativo. Tal mecanismo é capaz de observar o dispositivo adjacente e seus estados, e realizar modificações na estrutura de regras deste conforme estados ou entradas condizentes com as ações adaptativas existentes se manifestem.

São três as ações adaptativas possíveis, de consulta, inclusão e exclusão, e são capazes de observar e realizar mudanças na estrutura de regras do dispositivo adjacente, de modo que este passe a trabalhar de outra maneira depois da execução de uma ação adaptativa. Vale notar que as regras das ações adaptativas devem estar codificadas no mecanismo adaptativo, para que este saiba quando usá-las e para que efeito.

Há aplicação e formalismo de Adaptatividade para diversos dispositivos, tais como o Autômato Adaptativo, tabelas de estado e de decisão, gramáticas, cadeia de Markov e árvore de decisão, o que abre possibilidades acerca da melhor estratégia a ser adotada.

VI. Possibilidades de aplicação da adaptatividade

As aplicações da adaptatividade em data mining são muitas e a área é uma das que mais se beneficia da possibilidade de modificação adaptativa *on-the-fly* durante seu funcionamento. Passando pela obtenção de dados a partir de requisições web, *parsing* das árvores de dados, busca e extração da informação desejada, até o controle de acesso às fontes, ser capaz de modificar as regras pelas quais um dispositivo funciona conforme seu estado e entradas é muito valioso para quem trabalha com extração de informação, quer seja estruturada, semi ou não-estruturada.

Devido à aplicação e natureza dos dados judiciais extraídos pelos robôs atualmente, todas as informações coletadas precisam ser o mais específicas e corretas possível, o que inviabiliza num primeiro momento robôs adaptativos capazes de modificar-se para continuar extraindo dados mesmo com mudanças no sistema-alvo, já que assim passa-se a validação dos dados de dentro dos robôs para dentro da base de dados onde os processos são armazenados, com maior risco de envenenamento e empobrecimento da qualidade dos dados já obtidos. Ainda assim, espera-se que no futuro e com maior tempo para pesquisa na área possa-se aplicar algum desenvolvimento nesse sentido.

Já outra área tanto ou mais crítica no funcionamento da infraestrutura, e com menor risco em relação à diminuição da qualidade dos dados extraídos, é o controle de funcionamento dos diversos robôs que rodam durante 24 horas todo dia. As dificuldades no controle manual do funcionamento dos robôs, conforme descrito anteriormente, tem tomado uma proporção muito grande, e um sistema adaptativo, capaz de ter suas regras de funcionamento modificadas a partir das características de cada robô e sistema-alvo, podem ter um grande impacto positivo em termos de cobertura de extração, diminuição de custos e de tempo empregado na manutenção do funcionamento deste.

Atualmente cada robô tem codificada uma gama de informações a respeito de seu sistema-alvo, como tratamento de indisponibilidade, uso de *login*, bloqueio por *captcha* e mudanças de estratégia para acesso à informação, enquanto o *script* de controle dos robôs tem outros dados complementares já obtidos e tratados ao longo do tempo com respeito principalmente a horários de acesso e volume de acesso paralelo permitido. No entanto, não há um sistema que tome conhecimento dessas informações todas e as transforme em tomada de decisão automática. A geração de uma árvore ou tabela de decisão universal que leve em conta todos esses dados de maneira única é complexa e foge ao controle depois de algumas iterações, devido principalmente às enormes possibilidades de interação entre as características presentes em cada sistema-alvo.

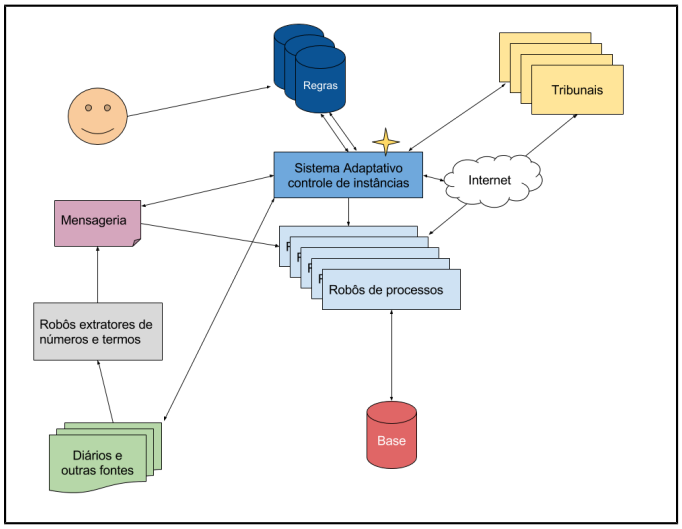

Fig. 3. Proposta de Sistema Adaptativo

A ideia de emprego da adaptatividade nesse ponto é a criação de um mecanismo adaptativo que atue sobre o subsistema de controle de instâncias e robôs e consiga analisar todos os dados sobre o alvo, presentes tanto nos robôs quanto no *script*, podendo então realizar ações adaptativas para modificar uma tabela de decisão padrão gerando um controle específico para cada robô. Além disso, com base no andamento das filas de processos, consiga ajustar de maneira iterativa essa tabela para um ajuste fino entre o solicitado e o atendido pelos robôs diariamente.

Outro ponto que se espera alcançar no desenvolvimento do mecanismo adaptativo é a capacidade deste de emitir alertas e avisos quando as regras e ações adaptativas existentes não forem capazes de garantir o funcionamento dos robôs e o consumo das filas em níveis aceitáveis, de modo que assim um operador possa realizar, ainda que manualmente, um segundo nível de adaptatividade, ajustando as ações adaptativas e a tabela de decisão para uma melhor conformidade com uma situação não antecipada.

VII. Futuros desenvolvimentos

A partir da decisão do caminho a ser seguido, os próximos passos agora serão a geração da tabela de decisão não-adaptativa do dispositivo adjacente, seguida da compilação em tabela das características de cada sistema-alvo para então codificação das ações adaptativas possíveis e os efeitos pretendidos na tabela de decisão.

Para o desenvolvimento do mecanismo adaptativo usar-se-á das linguagens de programação Python ou Clojure. A escolha tem dois pontos: enquanto Python é uma linguagem fácil de se trabalhar, flexível e atualmente usada em todos os nossos sistemas em produção, Clojure é um dialeto LISP com vários novos conceitos, capaz de rodar em cima da JVM do Java (abrindo portas para sistemas enterprise existentes), além de toda a possibilidade que ser um LISP traz, como macros, sintaxe simplificada e poderosa (código é dado) e alto nível.

Como base de regras para as ações adaptativas espera-se fazer uso de JSON ou uma base SQLite, facilitando a edição das ações mesmo por um desenvolvedor que não tenha tido contato com os conceitos mais avançados, como a teoria formal de adaptatividade, autômatos adaptativos, dentre outros.

Com o sistema pronto, a etapa seguinte é o funcionamento em malha aberta (sem ação real) paralelamente ao sistema não-adaptativo existente, como forma de validação e benchmark, para depois entrar em produção.

VIII. Resultados esperados

Busca-se finalizar o desenvolvimento inicial do mecanismo adaptativo em fevereiro de 2017 para início dos testes, com entrada em produção em março do mesmo ano garantido o funcionamento satisfatório.

Como funcionamento satisfatório entende-se a redução, de forma automática, de pelo menos 60% no tamanho das filas acumuladas no sistema de mensageria semanalmente, com atendimento correto de pelo menos 95% das mensagens delas.

Espera-se também, ainda que não seja possível uma quantificação adiantada, a diminuição de alertas e avisos falso-positivos, uma melhora na qualidade desses avisos e uma menor necessidade de intervenção humana para o consumo completo das filas de tarefas sem aumento expressivo no custo mensal com infraestrutura.

Referências

[1] Site O Globo, “Conflagrado nos tribunais, Brasil tem um processo em andamento para cada dois habitantes”, http://oglobo.globo.com/brasil/conflagrado-nos-tribunais-brasil-tem-um-processo-em-andamento-para-cada-dois-habitantes-16822691, visitado em 15/12/2016.

[2] LEI Nº 11.419, DE 19 DE DEZEMBRO DE 2006 - Lei de Informatização do Processo Judicial. Cópia obtida em http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11419.htm, visitado em 15/12/2016.

[3] Site da empresa Digesto, http://digesto.com.br/servicos, visitado em 15/12/2016.

[4] Neto, João José. Autômatos em Engenharia de Computação - uma Visão Unificada. Primera Semana de Ciencia y Tecnología de la Sociedad Chotana de Ciencias y la Red Mundial de Científicos Peruanos Ciudad de Chota, Perú, Junio 22-27, 2003.