UNIVERSIDADE FEDERAL DO AMAZONAS

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

DEPARTAMENTO DE APOIO À PESQUISA

PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

INICIAÇÃO CIENTÍFICA EM MÉTODOS PROBABILÍSTICOS PARA EXTRAÇÃO DE DADOS DE FONTES TEXTUAIS

Bolsista: André Luiz Lopes Porto, CNPq

Manaus

2011

INICIAÇÃO CIENTÍFICA EM MÉTODOS PROBABILÍSTICOS PARA EXTRAÇÃO DE DADOS DE FONTES TEXTUAIS

UNIVERSIDADE FEDERAL DO AMAZONAS

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

DEPARTAMENTO DE APOIO À PESQUISA

PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

RELATÓRIO FINAL

PIB - E/0002/2010

INICIAÇÃO CIENTÍFICA EM MÉTODOS PROBABILLÍSTICOS PARA EXTRAÇÃO DE DADOS DE FONTES TEXTUAIS

Bolsista: André Luiz Lopes Porto, CNPq

Orientador: Prof. Dr. Altigran Soares da Silva

Manaus

2011

Resumo:

Esse projeto tem como objetivo criar uma ferramenta para apoiar usuários finais em tarefas de extração de dados disponíveis em conteúdo textual, baseando-se para isso no método ONDUX. Este método foi desenvolvido pelo grupo de Banco de Dados e Recuperação da Informação (BDRI) da Universidade Federal do Amazonas (UFAM), tendo sido aplicado com sucesso em várias aplicações prototipais. No entanto, a sua ampla utilização em aplicações reais é atualmente restrita justamente pela ausência de recursos de interface que auxiliem o usuário.

A ferramenta foi implementada em linguagem Java, e já possui uma interface gráfica para que o usuário final possa utilizá-la. A interface é simples e objetiva, nela é possível observar todos os passos envolvidos no processo de extração de forma detalhada, deixando o usuário mais perto dos processos de extração. Além disso, é possível exportar as informações para dados estruturados como: XML e CSV, facilitando a manipulação do resultado da extração pelo usuário final.

SUMÁRIO

1. Introdução

1.1 Delineamentos da pesquisa

A abundância de documentos de texto contendo registros de dados na forma de texto contínuo tais como descrições de produtos, citações bibliográficas, endereços postais, anúncios classificados, etc., tem atraído uma série de esforços para extrair automaticamente os valores que compõe estes registros, segmentando o texto e classificando seus atributos. Por isso, faz se necessário: (1) a implementação de métodos capazes de extrair informações textuais e (2) a criação de ferramentas capazes de auxiliar o usuário durante o processo de extração.

Neste projeto exploramos a utilização de um método probabilístico para realização destas tarefas e desenvolvemos uma ferramenta para apoiar usuários finais em tarefas de extração de informação automática utilizando este método.

1.2 Hipótese

Para a extração correta desses registros dispostos em forma de texto contínuo, são necessários métodos capazes de extrair esses dados e classificá-los. Nesse cenário, o ONDUX (CORTEZ *et al,* 2010). é um método de extração de informação de fontes textuais que permite a realização de tarefas de extração em registros não estruturados através de segmentação de texto. Ao contrário dos métodos probabilísticos existentes na literatura (BORKAR *et al*, 2001; CORTEZ *et al*, 2007, CORTEZ *et al*, 2009,FREITAG;MCCALLUM, 2000), os quais necessitam de um conjunto de treinamento criado por um usuário especialista de cada aplicação, no ONDUX o processo de extração se baseia unicamente em conjuntos de

dados pertencentes a cada domínio, chamados de *bases de conhecimento*[1]. Estes conjuntos dados são facilmente encontrados na Web.

O presente projeto de iniciação científica visa à criação de uma ferramenta de extração de informação utilizando o método ONDUX para classificar corretamente dados dispostos em forma de textos contínuos.

2. Revisão bibliográfica

A Extração de Informação através da Segmentação de Texto (EIST) se aplica aos casos em que os valores de dados de interesse são organizados em registros implícitos e semi-estruturados disponíveis em fontes textuais (por exemplo, endereços postais, informação bibliográfica, os anúncios). (CORTEZ *et al*, 2010).

Existem dois tipos de extração de informação por segmentação de texto. Uma utiliza-se de técnicas de aprendizagem de máquina, na qual é necessário o usuário para executar o treinamento. Na outra, o treino é realizado por uma base de dados pré-existentes, eliminando o trabalho do usuário de realizar o treino manualmente.

Um dos métodos encontrados na literatura baseia-se na utilização de Hidden Markov Models (HMM). O primeiro trabalho que segue esta abordagem foi proposto por Freitag e McCallum (2006) e consistiu na geração independente do (HMM) para reconhecer os valores de cada atributo.

Posteriormente surgiram modelos de extração baseados no Conditional Random Fields (CRF), os quais foram propostos como uma alternativa ao HMM para tarefas de EIST. Dois trabalhos destacam-se na utilização do CRF: PENG;MCCALLUM, 2006; e MANSURI;SARAWAGI, 2006.

---

[1] Conjunto de palavras que pertencem a um determinado atributo.

Mais recentemente foi proposto o ONDUX (CORTEZ *et al*., 2010), um método de EIST que utiliza uma base de dados preexistente, conhecida como base de conhecimento, que vai auxiliar no reconhecimento dos atributos e classificação do texto de entrada. O ONDUX é um método para EIST que é considerado o estado-da-arte em extração de informação. Atualmente o método se encontra implementado através de um conjunto de scripts em linguagem PERL, tornando-o restrito a usuários avançados. A criação de um aplicativo facilitará o contato do usuário comum com os métodos de extração da informação.

A criação de uma ferramenta com uma interface gráfica que auxilie o usuário final na extração de vários tipos de dados dispostos em forma textual é importante e motivada pela necessidade de gerar esses dados em algum formato estruturado como bancos de dados relacionais ou XML, para que eles possam continuar sendo consultados, processados e analisados.

Para auxiliar na criação desta interface foi utilizado o Jgraph e IDE`s JAVA como o Eclipse e o Netbeans. Segundo Geer (2002), O Eclipse, além de proporcionar uma IDE, automatiza inúmeras funções que os desenvolvedores de código precisam, caso contrário, teriam que criar a "mão".

JGraph, é um sistema baseado em Java para desenhar gráficos e para a execução de algoritmos de grafos. Uma série de algoritmos bem conhecidos são fornecidos, incluindo os testes de planaridade e desenhos de grafos planares em uma grade (BAGGA;HEINZ, 2010). Para solucionar possíveis problemas no desenvolvimento, foi utilizado Padrões de Projeto, segundo Dantas *et al*(2002) um padrão resolve um problema recorrente, em um determinado contexto, fornecendo uma solução que comprovadamente funcione, além de informar os resultados e compromissos da sua aplicação, e subsídios para que seja possível adaptar esta solução a uma variante do problema.

O Padrão Model View Controller é muito utilizado para resolver problemas voltados a implementação de interfaces de usuário. O MVC separa o problema em três tipos de objeto, Modelo que é o objeto da aplicação, Visão que é apresentação na tela e o Controlador que define a maneira como a interface reage às entradas de usuário.

3. Métodos Utilizados

- Pesquisa bibliográfica.
    - Método ONDUX.
        - Estudo do método de extração textual ONDUX de acordo com o artigo publicado para entender o seu funcionamento.
    - User Interface em linguagem JAVA.
        - Pesquisa sobre UI (User Interface) visto que foi necessário referencial teórico para a criação da ferramenta.
    - Padrões de projeto.
        - Pesquisa e estudo do padrão de arquitetura de software Model View Controller (MVC) .
    - Biblioteca Jgraph.
        - Pesquisa e estudo da biblioteca Jgraph para criação gráfica de uma das etapas do método Ondux.

- Implementação das técnicas pesquisadas.
    - Implementaçao do método Ondux na linguagem java.
    - Criação da UI(User Interface).
    - Criação do representador gráfico(PSM) utilizando a biblioteca Jgraph.
    - Criação de Website para divulgação de dados da ferramenta.

- Execução de testes e avaliação da ferramenta.
    - Após a criação da ferramenta é necessário iniciar testes para verificar se a mesma está funcionando corretamente.

4. Resultados e discussões

O método ONDUX (CORTEZ *et al,2010),* envolve três etapas principais, *Blocking*, *Matching* e *Reinforcement*. Essas etapas representam a segmentação do texto e posteriormente sua classificação. O método é baseado em um modelo probabilístico que resulta na probabilidade de um certo termo representar um atributo, avaliando também a posição e sequência dos dados de entrada.

A ferramenta necessita de duas entradas, que são carregadas antes da execução da classificação, uma representa a entrada do usuário a qual ele deseja efetuar a classificação e a outra referente a uma base de conhecimento.

Uma base de conhecimento K pode ser definida como um conjunto de pares tal que $K = \{(<a_{1,} O1>)...(<a_{n,} On>)\}$,onde a_i representa atributos distintos e O_i representa um conjunto de strings $\{O_{i,1},...,O_{i,n_i}\}$ chamados de ocorrências. Intuitivamente, O_i representa um conjunto de strings que representam possíveis valores referentes a cada atributo $a_{i.}$

Exemplo de uma base de Conhecimento: cada linha representa uma ocorrência que compõe a base de conhecimento.

```
<kb>
      <name> Universidade Federal do Amazonas </name>
      <name> Universidade Estadual do Amazonas </name>
      <street> Rua General Rodrigo Otávio</street>
      <street> Avenida Darcy Vargas </street>
      <city> Manaus </city>
      <city> Parintins </city>
      <phone> (092) 1121-2222</phone>
      <phone> (55) 92 1111-2222</phone>
</kb>
```

Quadro 1 – Exemplo de Base de Conhecimento

Para o desenvolvimento da ferramenta, foi utilizado a linguagem JAVA por ser uma linguagem de fácil usabilidade e capaz de dar um suporte maior para a criação de uma interface mais robusta e auxílio na utilização de bibliotecas disponíveis na internet. A implementação seguiu criteriosamente os algoritmos apresentados no artigo publicado, dando fidelidade ao método de extração. Modelos de padrões de projeto foram utilizado, como o MVC que auxiliou no desenvolvimento da interface gráfica

4.1 Implementação do ONDUX na Linguagem Java

Atualmente, todo o método se encontra implementado na linguagem. O início da implementação foi dado pela etapa de *Blocking*, etapa a qual é realizada a segmentação do texto em pequenos blocos simplesmente separados por espaços em branco encontrados entre as palavras a partir dos dados de entrada. Posteriormente nessa etapa os blocos são aperfeiçoados. Esses blocos que antes possuíam um termo podem ser modificados e mesclados com outros, o determinante desta união se dá na base de conhecimento, que irá verificar se os blocos possuem alguma relação entre os termos para que sejam unidos.

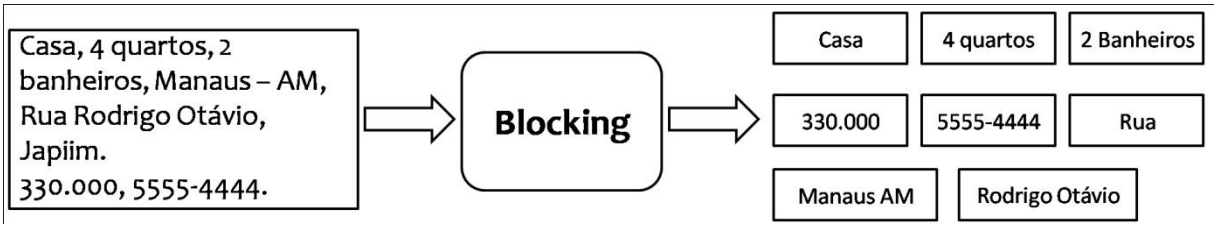

Figura 1. Etapa Blocking do método ONDUX.

Com a segmentação pronta, a etapa *Matching* entra em operação estimando probabilidades para classificar cada bloco colocando seu atributo correspondente, expressando melhor o impacto do termo na coleção que se quer classificar.

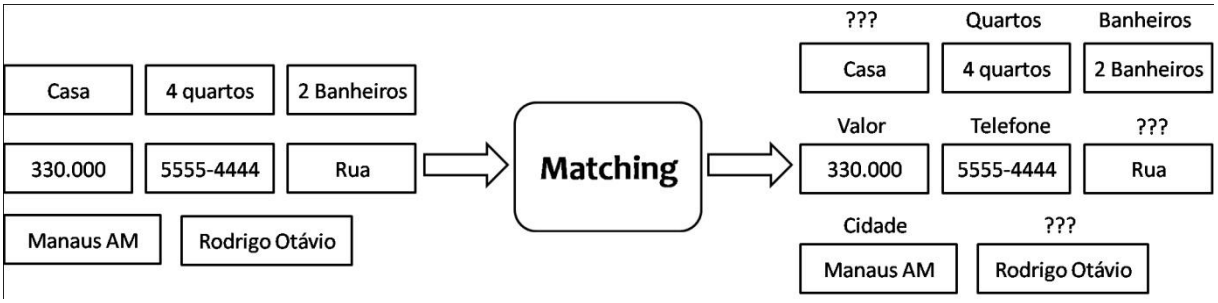

Figura 2 - Etapa Matching do método ONDUX.

A etapa final é chamada de *Reinforcement*, que significa reforço. Nesta etapa um modelo probabilístico semi-Markoviano chamado PSM (Positioning and Sequence Model) é usado para corrigir problemas na classificação a partir de dados de seqüenciamento e posição de cada estrutura de entrada da ferramenta. Com isso a probabilidade de ter um termo corretamente rotulado aumentará.

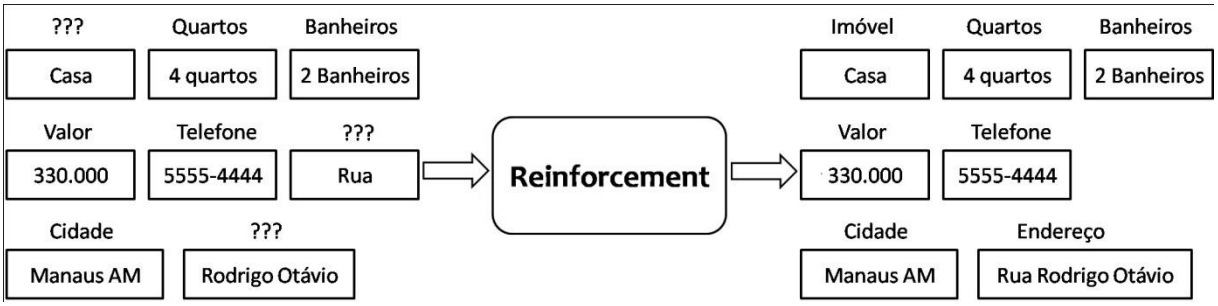

Figura 3 - Etapa Reinforcement do método ONDUX.

4.2 Gerador do Grafo PSM

Como mencionado, o método ONDUX possui um etapa de reforço que avalia a posição e a sequência em que os termos estão dispostos. Para representar graficamente essas posições foi desenvolvido um gerador de grafo que mostra os atributos e suas respectivas probabilidades desse atributo aparecer antes ou depois de outros atributos. Esse gerador torna o usuário mais próximo dos processos computacionais e pode ser utilizado para verificar uma semi-estrutura das fontes textuais de entrada.

Para criação desse gerador PSM foi utilizado a biblioteca Jgraph, disponível na Web[2], que auxiliou na criação de vértices e arestas do grafo, podendo atribuir nomes e pesos de uma forma dinâmica e de fácil manipulação visando a facilidade do programador em desenvolver o gerador.

Os dados dos vértices, pesos e arestas são gerados pela função *Reinforcement*. O gerador interpreta essas informações e desenha graficamente na tela do computador representando o PSM.

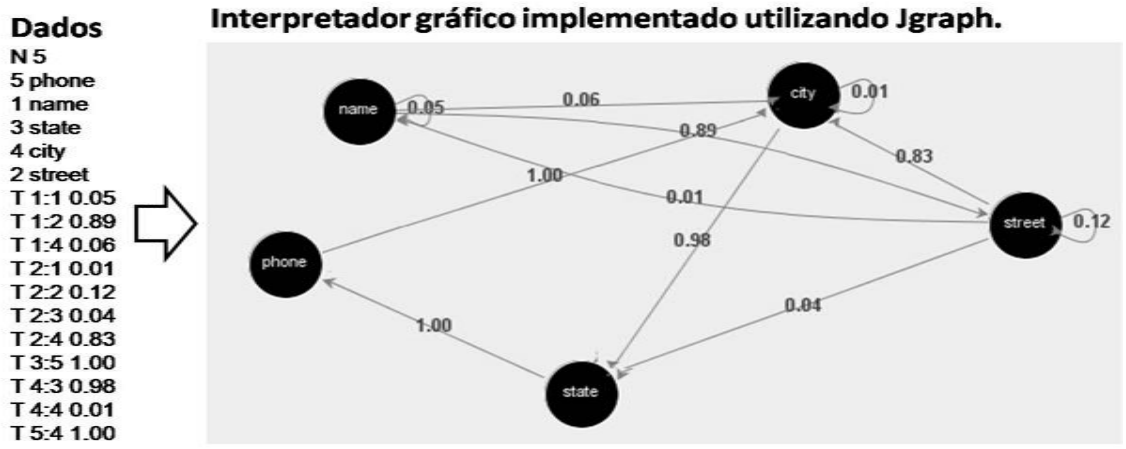

Figura 4 - Interpretação do gerador de grafo.

4.3 Desenvolvimento de uma versão *stand-alone* da ferramenta utilizando uma GUI *(Graphical User Interface*).

Para a criação da *User Interface* foi utilizado uma IDE JAVA chamada NetBeans, que auxiliou no desenho e na adição das funcionalidades da ferramenta, criando a interface entre programa implementado e o usuário.

[2] http://www.jgraph.org/

A interface tem como objetivo transformar o método ONDUX em uma ferramenta que irá auxiliar o usuário final em tarefas de extração automática de informações de uma forma fácil e intuitiva.

O objetivo principal do projeto se foca nesse tópico. Para isso é necessário criar uma versão stand-alone que representa um programa de computador com janelas e botões que suporta o carregamento da base de conhecimento e do texto que se deseja classificar.

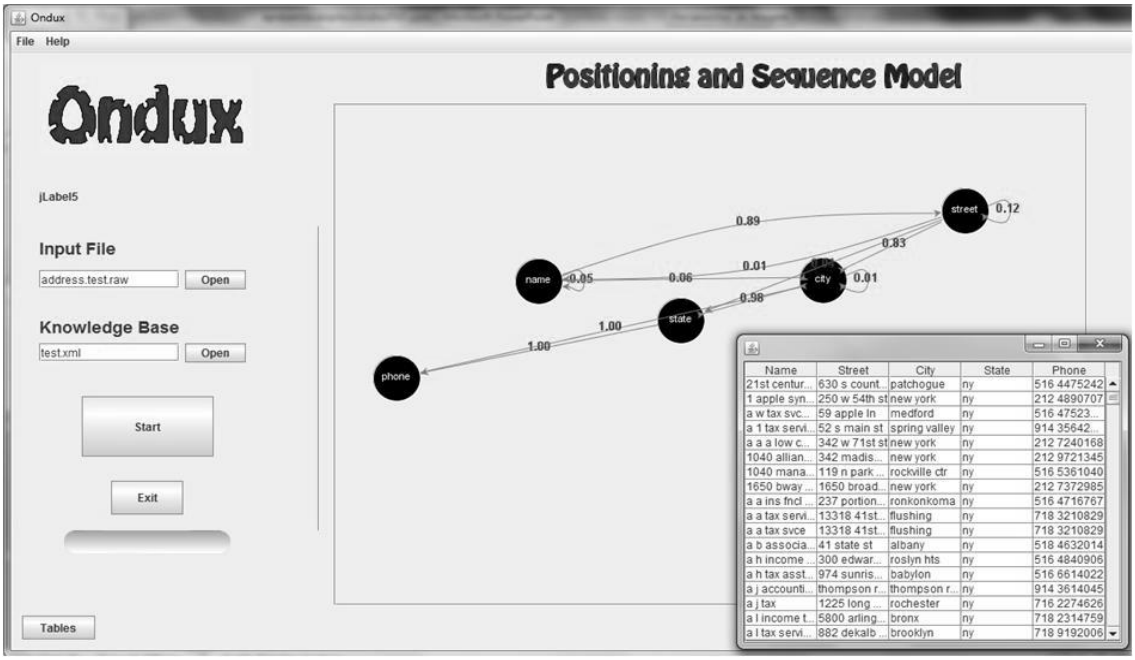

Figura 5- Interface gráfica inicial da ferramenta.

4.4 Aprimoramentos da ferramenta

Foi criada uma segunda versão final da ferramenta que tem como objetivo principal, exibir as etapas importantes do método ONDUX, visando necessidades didáticas e científicas. Nessa versão o usuário poderá acompanhar cada etapa do método separadamente com seus respectivos dados. Abas foram criadas para cada etapa: *Blocking, Matching e Reinforcement*. Com isso o usuário será capaz de observar as etapas de execução do programa de forma mais minuciosa, verificando os blocos que foram classificados ou não diretamente na janela da

ferramenta. Cada bloco possui uma classificação e atributos que não foram classificados são tachados como "*UN*"(*Unmatched*).

Nessa versão a interface foi desenvolvida a partir do padrão de projeto MVC, com isso o usuário é capaz de executar o método e ao mesmo tempo verificar os dados que estão sendo processados mesmo antes de acabar o método. Esse padrão de projeto solucionou um problema de espera do término da execução, na versão anterior da ferramenta, o usuário necessitava esperar todo o processamento do método, visto que a arquitetura da aplicação desenvolvida não utilizava threads para renderização da janela. A janela ficava travada até que todo método fosse executado. Com a utilização dos conceitos de MVC e manipulação de threads foi possível separar o processamento da interface e resolver esse problema.

Além disso, foram criadas mais duas funcionalidades muito importantes, responsáveis pela exportação do resultado de classificação para dados estruturados, como o XML ou CSV a fim de que o usuário possa manipular essas informações de forma mais eficiente e objetiva.

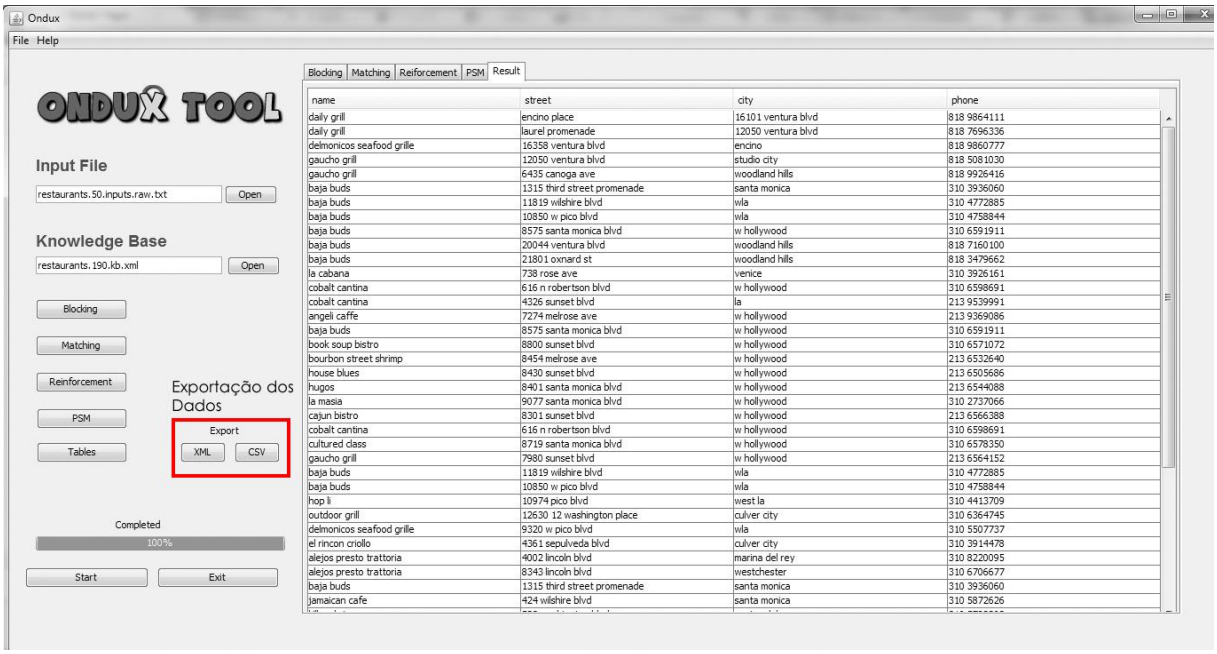

Figura 6 - Versão final da ferramenta

4.5 Criação de site para divulgação de trabalhos

Com objetivo de disponibilizar mais informações sobre o projeto e a ferramenta, foi criado um website. O Site está disponível em *http://www.gtiexperimentos.com.br/Ondux* e possui vários arquivos para serem executados com a ferramenta, esses arquivos possuem uma entrada que se deseja classificar e várias bases de conhecimento.

4.6 Realização de testes e avaliação da ferramenta

Para apresentar uma ferramenta estável computacionalmente, foi realizado vários testes com distintas bases de conhecimentos e entradas de usuários, muitas delas são as utilizadas em eventos científicos e usadas como avaliação de vários modelos de extração de informação. Esses testes visam verificar erros e buscar soluções para os que possívelmente poderiam aparecer. Todos testes foram executados e os erros encontrados, foram solucinados. As seguintes bases foram utilizadas nos testes:

Base de Conhecimento	Conteúdo das bases	Número de Entradas	Número de Atributos
BigBook	Endereços	2000	5
CORA	Referências Bibliográficas	500	3 até 7
Web ADS	Classificados	500	5 até 8
Receitas	Receitas de cozinha	500	3
Produtos	Oferta de produtos	10.000	3

Quadro 2 – Bases utilizadas

5. Conclusões

Indubitavelmente sabe-se que muitas pesquisas científicas acabam não atingindo todas esferas de usuários, muitas acabam ficando somente no âmbito acadêmico, com profissionais

da área, ou são utilizadas em outras pesquisas. Deste modo, criar uma ferramenta destinada a usuários finais torna o conceito científico mais forte, pois auxilia a sociedade como um todo resolvendo seus problemas a partir de soluções computacionais. Neste projeto foi criada uma ferramenta de extração automática de informação utilizando um modelo conhecido na literatura como ONDUX que foi desenvolvido na Universidade Federal do Amazonas mais precisamente no laboratório de Banco de Dados e Recuperação da Informação. A ferramenta recebe vários termos semi-estruturados que necessitam ser classificados e uma base de conhecimento, com isso, a ferramenta extrai dados intrínsecos da entrada e estabelece uma relação entre os termos, classificando-os, com algum respectivo atributo.

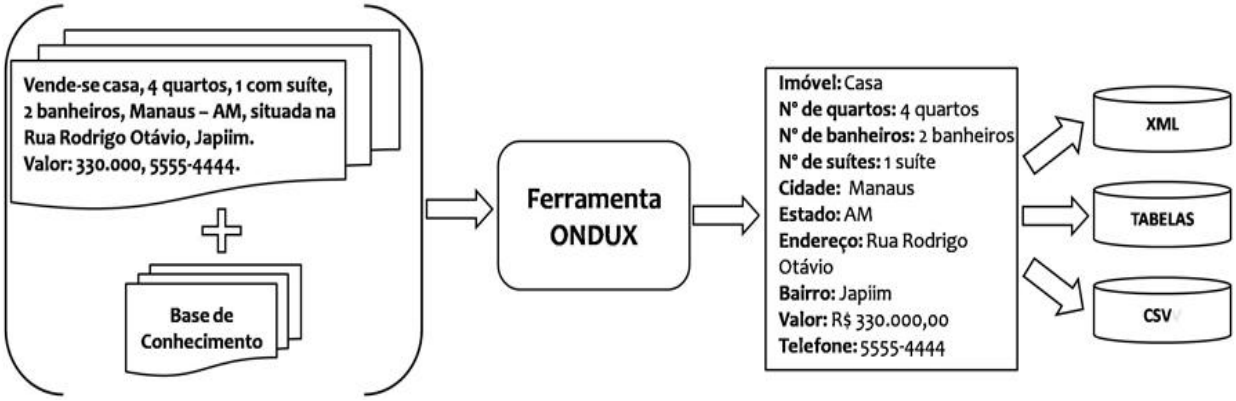

Figura 7. Processo de extração utilizando a ferramenta.

A iniciação científica na área de Recuperação da Informação se torna importante ao mesmo passo em que a quantidade de informação disponível, principalmente na Internet, tende a crescer. O usuário sempre está em busca informações, mas se depara com vários problemas visto que a maioria das informações ricas estão dispostas em textos contínuos provenientes de vários outros processos e em muitos casos essa informação se apresenta desordenada ou implícita, tornando o usuário incapaz ou dificultando a tarefa de extrair informações desses dados. Ferramentas como a criada nesse projeto auxiliam consideravelmente esse processo de extração, automatizando a tarefa e tornando o usuário um

intermediador e não executor do processo. A ferramenta abre oportunidades e incentiva a pesquisa de novo métodos para melhorar os resultados obtidos, vários trabalhos já estão sendo realizados a fim de encontrar um ponto ótimo na extração de informação utilizando novas técnicas. Futuramente pretende-se utilizar técnicas de Active Learning com objetivo de aumentar a precisão nos resultados, por enquanto, este trabalho apenas se objetivou na produção da ferramenta.

6. Referências Bibliográficas

BORKAR, V. R. ; DESHMUKH, K. ; SARAWAGI S. Automatic segmentation of text into structured records. Proc. of the ACM SIGMOD International Conference on Management of Data, pages 175-186, 2001.

CORTEZ, Eli; DA SILVA, Altigran Soares; GONÇALVES, M. .; DE MOURA, Edleno Silva Ondux: on demand unsupervised learning for information extraction. In SIGMOD ’10: Proceedings of the 2010 international conference on Management of data, pages 807–818

CORTEZ, Eli; DA SILVA, Altigran Soares ; GONCALVES, M.; MESQUITA, F.; DE MOURA, Edleno Silva. FLUX-CIM: flexible unsupervised extraction of citation metadata. Proc. of the 2007 conference on Digital libraries, pages 215-224, 2007.

CORTEZ, Eli; DA SILVA, Altigran Soares; GONÇALVES, M; MESQUITA, F.; DE MOURA, Edleno Silva. A flexible approach for extracting metadata from bibliographic citations. Journal of the American Society for Information Science and Technology, Online version, 2009.

DEITEL, H. M. ; DEITEL, P. J.; JAVA Como programar. 8 ed. São Paulo: Pearson education do Brasil, 2005.

DANTAS, Alexandre, et al. Suporte a Padrões no Projeto de Software. XVI Simpósio Brasileiro de Engenharia de Software, Gramado – RS, 2002. P.450 – 455.

FREITAG, D; MCCALLUM, A. Information extraction with hmm structures learned by stochastic optimization. In Proc. of the 17th National Conf. on Articial Intelligence and 12th Conf. on Innovative Applications of Artificial Intelligence, pages 584-589

GAMMA,Erich, et al Padrões de Projeto – Soluções de software orientado a objetos, Reimpressão , BookMan, 2008.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. of the Eighteenth International Conference on Machine Learning, pages 282–289, 2001.

MANSURI, I. R.; SARAWAGI, S. Integrating unstructured data into relational databases. In Proc. of the International Conference on Data Engineering, page 29. 2006.

MENDES, Douglas Rocha. Programação Java com Ênfase em Orientação a Objetos. 1 ed. São Paulo: Novatec, 2009. pp. 456.

PENG, F. ; MCCALLUM, A. Information extraction from research papers using conditional random fields. Information Processing Management, 42(4):963-979.

7. Cronograma Executado

	Descrição	Ago 2010	Set	Out	Nov	Dez	Jan 2011	Fev	Mar	Abr	Mai	Jun	Jul
1	Estudo do método de extração textual ONDUX.	X	X										
2	Implementação das técnicas estudadas em JAVA.			X	X								
3	Estudos dos requisitos de interface gráfica necessários à utilização das técnicas implementadas.					X	X						
4	Implementação dos recursos de interface estudados.							X	X				
5	Realização de testes e avaliação da ferramenta.									X			
6	Elaboração do Resumo e Relatório Final (atividade obrigatória).										X	X	
7	Preparação da Apresentação Final para o Congresso.												X