Técnicas de Extração de Informação para Avaliação da Qualidade de Páginas Web com o Uso de Ontologias

Daniel Lichtnow[1,2]

José Palazzo M. de Oliveira[1]

José Valdeni de Lima[1]

Mario Lemes Proença Jr.[3]

Rodolfo Miranda de Barros[3]

Mário Henrique A. C. Adaniya[3]

Resumo: A qualidade dos conteúdos das páginas *Web* pode ser determinada parcialmente através de indicadores como autoria da página, existência de referências as fontes e reputação do responsável pela publicação. Este artigo discute a aplicação de técnicas de extração de informação na identificação de indicadores de qualidade, especificamente autoria. Ao contrário de outras técnicas de extração, as técnicas desenvolvidas neste trabalho não utilizam a estrutura das páginas. Neste sentido, o artigo apresenta os resultados iniciais do trabalho, aponta algum dos problemas envolvidos e identifica técnicas que podem ser úteis para continuidade do trabalho. O objetivo final do trabalho é criar uma ferramenta que possibilite avaliar a qualidade de sites com conteúdo relacionado à área de saúde. Assim, os resultados do processo de extração são

[1] Instituto de Informática, Universidade Federal do Rio Grande do Sul, Caixa Postal 15.064
{palazzo, valdeni, dlichtnow@inf.ufrgs.br}
[2] Centro Politécnico, Universidade Católica de Pelotas - Rua Felix da Cunha 412, Pelotas, RS, Brasil
{lichtnow@ucpel.tche.br}
[3] Dept. de Computação, Universidade Estadual de Londrina, Caixa Postal 6001
{proenca, rodolfo@uel.br; mhadaniya@gmail.com}

utilizados para popular uma ontologia onde estão definidos os critérios de qualidade para as páginas *Web*.

Abstract: The content quality of Web pages can be determined partially by indicators such as authorship, presence of references and publisher reputation. This paper discusses the application of information extraction techniques on the identification of quality indicators, specifically authorship. Unlike other extraction techniques, the techniques of our work, try to make the extraction without consider the structure of Web pages. The final goal of our work is to create tools to assess the quality of web pages that have topics related to health. In this sense, the goal is to use the results of extraction to populate an ontology where are defined quality criteria for web pages.

1 Introdução

A crescente preocupação com a qualidade do material disponível na *Web* tem levado a definição de critérios e indicadores de qualidade para *sites* por parte de algumas organizações[4]. Uma das dificuldades de avaliação da qualidade reside no fato de que o processo de avaliação não é automatizado, assim cada *site* ou página precisa ser analisado manualmente pelos membros de organizações que concedem selos de qualidade para os que atendam critérios específicos ou mesmo pelos usuários, que ao acessar uma página, julgam seu conteúdo, nem sempre usando critérios adequados. Estes problemas motivaram a criação dos Projetos SALUS CYTED e Prosul AvalSaúde, CNPq, que envolvem universidades da América Latina e Europa (Portugal e Espanha) e visam construir ferramentas que permitam avaliar a qualidade das páginas *Web* relacionadas a temas da área de saúde.

Dentro destes projetos, está sendo feita a construção de uma ontologia que tem por objetivo representar os critérios de qualidade e automatizar parte da avaliação da qualidade do conteúdo de páginas *Web*.

[4] http://www.hon.ch

Um problema para automatizar a avaliação está relacionado à dificuldade de extrair das páginas *Web* os indicadores de qualidade que estão representados na ontologia. Dentre os indicadores de qualidade, considerados relevantes, está à indicação da autoria do conteúdo.

A partir disto, neste artigo é apresentada uma proposta para realizar o processo de identificação dos nomes dos autores mencionados nas páginas *Web*. O resultado da extração é usado criar instâncias na ontologia que representa os critérios de qualidade.

Cabe salientar que uma vez que o projeto SALUS busca avaliar a qualidade de qualquer página *Web,* é impossível determinar previamente qual a estrutura destas páginas, tornando muitas das técnicas de extração de informação, que tiram proveito da estrutura das páginas *Web,* inadequadas.

Neste sentido, o artigo apresenta alguns resultados iniciais do trabalho e técnicas que podem ser úteis para o seu prosseguimento.

Na seção 2 deste artigo são apresentados os trabalhos relacionados. Na seção 3 são apresentadas as técnicas utilizadas e também parte da ontologia de qualidade. Na seção 4 são apresentados os experimentos iniciais. Finalmente, a seção 5 apresenta as considerações finais.

2 Trabalhos Relacionados

O trabalho está relacionado à Extração da Informação – EI, que objetiva extrair informação estruturada a partir de fontes não estruturadas [11]. O resultado de um processo de extração pode ser utilizado para popular ontologias [9].

Muitas das técnicas utilizadas para realizar a extração de dados de páginas *Web,* levam em conta à estrutura das páginas. Assim, para aplicar estas técnicas, é necessário que a estrutura presente nas páginas utilizadas como fonte esteja claramente identificada. Quando não, isto é, quando as fontes de dados são menos estruturadas são empregadas técnicas de Processamento de Linguagem Natural – PLN [6].

Para nosso projeto, a base de documentos adotadas são páginas *HTML*, que são caracterizados como documentos semi-estruturados, contendo uma tipificação através das *tags* que são utilizados para organizar o conteúdo e a visualização do próprio documento. Nossa base de documentos adotada é a Web, especialmente os sites voltados para áreas de saúde.

Em função dos requisitos dos projetos SALUS e ProSul (ver seção 3.2), são mais relevantes para o trabalho, as técnicas de EI que procuram realizar o processo de extração usando pequenos conjuntos de padrões domínio-independentes. Um exemplo é apresentado em [2], onde o sistema de extração procura identificar instâncias de uma entidade a partir da ocorrência de padrões como "*X é uma Y*" (*X* é uma instância de uma classe *Y)* Assim, o padrão *"* é um autor"* retornaria, em princípio, instâncias da classe autores.

No que se refere ao uso de técnicas de extração de informação para avaliação de sites da área de saúde, em [4] é descrita a ferramenta AQUA, que automatiza parte do processo de avaliação de *sites* por organizações que concedem selos de qualidade para *sites* (presença dos autores é um critério considerado). A ferramenta AQUA aprimora seus resultados como base no *feedback* fornecido pelo usuário. Este procedimento, não é adequado aos requisitos dos projetos SALUS e ProSul (seção 3.2), pois espera-se a geração de indicadores que orientem o público leigo (usuários Web sem conhecimento na área de saúde) nem sempre apto a fornecer o *feedback* adequado.

Já em [5], o objetivo, como no presente trabalho, é identificar o nome do autor em uma página *Web*, uma área importante dentro da EI, *Named Entity Recognition (NER)*. No trabalho, o conteúdo das páginas (em japonês) é analisado com um *parser*, sendo identificados os nomes dos possíveis autores. Dentre estes, um é identificado como autor (o trabalho considera que o nome do autor está próximo do conteúdo principal).

Named Entity Recognition - NER normalmente envolve o reconhecimento de nomes próprios relacionados a pessoas, localizações e organizações. Porém, existem outros trabalhos que procuram identificar outros tipos de entidades relacionados a necessidades específicas (filme, ator, livro, autor etc.) [3].

Em [12], o objetivo é uma arquitetura para *NER* utilizando *Link Grammar* e o algoritmo *Basilisk*. *Link Grammar* é um analisador sintático baseado em uma teoria original da sintaxe Inglesa. Dada uma sentença, uma estrutura sintática é construída composta por um conjunto de links conectando pares de palavras. O algoritmo *Basilisk* é um algoritmo de aprendizado léxico semântico com resultados de alta qualidade. Dado um conjunto de corpus de documento, ele hipotetiza a classe semântica de uma palavra. Com isto, os autores apresentam uma arquitetura para *NER* aplicada em relatórios sobre saúde.

Em [14], o autor determina uma hierarquia de domínio entre as páginas, de acordo com a similaridade dos conteúdos, criando uma rede interligada por conteúdos similares. Mas o dicionário e padrões para regras de associação dependem do cenário da informação que se deseja extrair. O reconhecimento de entidades nomeadas é baseado nas regras criadas entre os domínios e cenários estabelecidos anteriormente. Interessante notar que uma vez estabelecidos as regras e cenários, o nível de confiança de dada informação extraída pertencer a dado domínio no qual foi extraído é alto.

Trabalhando com o conceito de hierarquia, em [8], o autor propõem um modelo de extração hierárquica baseada em Modelos *Hidden Markov*. Átomos de elementos de informação e informações de itens compostos são extraídos de documentos HTML, através de métodos *bottom-up* gerando uma árvore. Através dos caminhos possíveis de se percorrer a árvore, os modelos de extração são criados.

Em contraste com os projetos que utilizam a Web como um banco de dados, como um dicionário ou de outra maneira, em [10], podemos observar que quando aplicado para um determinado conjunto de documentos previamente conhecidos, e possuindo um

conhecimento maior do domínio e cenário trabalhado, a arquitetura e ferramentas agregam valores em cenários reais, no caso, o cenário real de um hospital onde um médico ou outro funcionário autorizado consegue extrair informações dos documentos coletados do paciente, de maneira eficiente.

3 Abordagem Proposta

No trabalho, são definidas e implementadas técnicas que auxiliam no processo de identificação dos autores do conteúdo de uma página.

Cabe ressaltar que está se buscando identificar o nome do autor presente/declarado na página. *Parsers* são utilizados apenas no tratamento das páginas HTML, eliminando assim a marcação e obtendo o conteúdo. Sobre o conteúdo, um *parser* construído apenas com expressões regulares, e posteriormente a *Web*, que é usada como um dicionário. Importante destacar a diferença entre a utilização de um dicionário previamente definido com a utilização da abordagem de utilização da Web como um dicionário. Quando definimos previamente o dicionário, o conteúdo é fixo. Geralmente o dicionário é utilizado para eliminar expressões extraídas pela ferramenta, quando dada expressão tem como componente uma parte de algum elemento do dicionário, não sendo considerado uma expressão válida, dependendo do contexto. A idéia de utilizar a Web como um dicionário, é também estrita para um contexto, mas é maleável o suficiente para trabalhar com outros domínios.

Como a informação sobre a autoria será utilizada na avaliação da qualidade da página, uma parte da ontologia definida para avaliação de qualidade é apresentada a seguir. São também apresentados os requisitos relacionados às técnicas de EI.

3.1 Ontologia de Qualidade e o Processo de Avaliação

A ontologia para avaliar a qualidade de Páginas web é definida com a utilização de recursos da *Web* Semântica. Os critérios de qualidade são baseados parcialmente no *HONCode*[5]. Por questões de espaço, no trabalho apenas parte da ontologia é apresentada. Na ontologia, seguindo [1], são definidas 3 classes dentro das quais uma página *Web* poderá estar classificada: *Satisfactory*, *Marginal* e *Unsatisfactory*. Os critérios usados na ontologia para classificar uma página *Web* são apresentados, de forma sucinta, em (1) usando o estilo de apresentação do Protégé[6].

$$\begin{gathered}\textbf{Satisfactory\_Intrinsic\_Quality} \\ ((\forall \text{ hasAuthor Expert\_Author}) \vee (\forall \text{ qualityReferences High})) \wedge \\ (\forall \text{ hasOwner Recognized})\end{gathered} \quad (1)$$

[5] http://www.hon.ch

[6] http://protege.stanford.edu/

Marginal_Intrinsic_Quality
(((∀ hasAuthor Expert_Author) ∨ (∀ qualityReferences High)) ∧
(∀ hasOwner Non_Recognized)) ∨
(((has Author Non_Expert_Author) ∧ (∀ qualityReferences Low))
∧ (∀ hasOwner Recognized))
Unsatisfactory_Intrinsic_Quality
((∀ hasAuthor Non_Expert_Author) ∧ (∀ qualityReferences Low))
∧ (∀ hasOwner Non_Recognized)

Seguindo o exposto em (1), uma página *Web* tem maior qualidade quando pertence a uma organização reconhecida (∀*hasOwner Recognized*) ou tem seu conteúdo elaborado por especialistas (∀*hasAuthor Expert_Author*) ou aponta para boas referências (∀ *qualityReferences High*). Os critérios para classificar um autor como especialista estão fora do escopo deste artigo, sendo discutidos em outros trabalhos [7], mas a identificação do nome do autor é o primeiro passo.

3.2 Requisitos para Técnicas de Extração de Informação

Considerando os projetos (SALUS e ProSul) no qual está inserido este projeto, o objetivo é apresentar ao usuário (no momento que o usuário acessar a página) a avaliação feita com os critérios definidos na ontologia. Assim, o usuário acessa uma página *Web*, os dados necessários à avaliação são extraídos e inseridos na ontologia e a página é avaliada (classificada) a partir dos critérios presentes na ontologia. A partir disto, os seguintes requisitos relacionados às técnicas de extração são considerados:

- A tarefa de extração consiste basicamente em reconhecer entidades - *Named Entity Recognition and Classification* (subtarefa da EI que tem por objetivo reconhecer nomes de pessoas, locais e organizações, por exemplo);
- Devem ser usadas técnicas de extração que não estejam limitadas à extração de dados de páginas *Web*, cuja estrutura apresente certa padronização;
- Não se pode contar com o *feedback* dos usuários para aprimorar o processo.

3.3 Descrição das Técnicas de Extração de Informação Utilizadas

Basicamente, para realizar a identificação dos nomes dos autores do conteúdo de uma página *Web* são executados dois passos:

1. São identificadas, nas páginas *Web*, seqüências de termos que possam ser nomes de autores, usando expressões regulares;

2. São feitas consultas a *Web* utilizando um motor de busca, de forma a descartar algumas das seqüências de termos identificadas.

O uso de expressões regulares (passo 1) tende a identificar mais *Named Entities* do que aquelas que realmente existem. O passo 2 visa reduzir este número, aumentando a precisão. A consulta, realizada no passo 2, consiste na combinação da seqüência de termos, identificada com a expressão regular, com um padrão que permita confirmar se a seqüência de termos identificada como uma *Named Entity* é de fato uma *Named Entity* relacionada a uma classe específica (Pessoa/Autor).

Esta estratégia segue [2], onde para identificação de instâncias de uma classe na *Web* são utilizadas expressões em inglês que indicam a presença de instâncias de uma classe em textos presentes em páginas *Web*. Algumas das expressões (em inglês) utilizadas são mostradas em (2) juntamente com a tradução para Português (*NP - Noun Phrases* e *SN - Sintagmas Nominais*).

NP “and other” <class 1>	SN “e outra” <classe 1>	
NP “or other” <class 1>	SN “ou outra” <classe 1>	
<class 1> “such as” NPList	<classe 1> “tal como” SNList	(2)
“such” <class 1> “as” NPList	“tal” <class 1> “como” SNList	
NP “is a” <class 1>	SN “é um” <classe 1>	
NP “is the” <class 1>	SN “é o “ <classe 1>	

Em [2] o objetivo está em submeter estas expressões para um motor de busca de forma a extrair uma lista de nomes relacionados a uma classe a partir do conteúdo presente na *Web* e identificado e retornado pelo motor de busca. Deste modo, é submetida, por exemplo, a expressão “*is a city*” a um motor de busca, sendo identificados os sintagmas nominais que precedem a expressão e que são, em princípio, nomes de cidades(ex. *New York is a city)*.

No presente trabalho, como são relevantes apenas *Named Entities* da classe pessoas, foram consideradas outros padrões além dos definidos em [2]. Exemplos são *“Mr.”,* “*Sr.”, “papers by”, “escritos por”* seguidos de uma *Named Entity*. Ainda, médicos podem ser identificados de forma semelhante (“*John Smith, M. D.*”, “*João da Silva é médico*”, “*João da Silva é cardiologista*”), o que é relevante dado o domínio (saúde) dos projetos onde está inserido este trabalho.

4 Experimentos

A ferramenta foi construída usando *HTML Parser*[7]*,* JavaCC[8], considerando ainda o trabalho desenvolvido em [13]. Para as buscas em uma search engine, foram utilizadas:

[7] http://htmlparser.sourceforge.net/

[8] https://javacc.dev.java.net/

Google AJAX Search API[9] em conjunto com a biblioteca JSON[10]. As expressões regulares foram construídas manualmente, a partir de experimentos e o estudo de alguns exemplares de páginas. Para os experimentos iniciais foram usados um conjunto de páginas *Web,* retornadas por consultas realizadas usando o mecanismo de busca de uma biblioteca digital, relacionada à área de saúde[11]. Os resultados obtidos no primeiro experimento são mostrados na Tabela 1.

Para confirmar o fato, já esperado, de que ignorar as *tags* presentes em uma página *Web* e utilizar apenas uma expressão regular simples, faz com que a abrangência seja alta (um grande número de nomes recuperados) e a precisão baixa, foram criado dois cenários de implementação. No primeiro as *tags* com a *string* "*author*" foram consideradas para a busca de nome de autores, e no segundo cenário, nenhuma *tag* foi levada em consideração.

Como resultado, a primeira abordagem a precisão é alta, mas dependente da estrutura das páginas, enquanto que na segunda abordagem, a precisão diminui. Isto não significa que os nomes não são extraídos, apenas juntamente com nomes válidos, outras informações também são extraídas, como nome de lugares, possíveis nomes de doenças e informações incorretas como início de parágrafo.

Cabe salientar que, conforme destacado na seção 3.2, considerar a estrutura específica de uma página *Web*, não é uma abordagem adequada dados os requisitos do trabalho.

Tabela 1. Resultados do Primeiro Experimento

Legenda	Total
Nome presentes	273
Nomes identificados pelo programas	417
Nomes corretamente identificados	260
Precisão – Precision	62%
Abrangência – Recall	95%

Em um segundo experimento, foram utilizados alguns dados usados no primeiro experimento (Tabela 1), juntamente com as expressões e estratégia apresentadas na seção 3.3 (passo 2). Assim, se a seqüência de termos identificada como um possível nome de autor por meio da expressão regular, combinada com algum dos padrões mostrados em (2), for encontrada em alguma página em uma pesquisa feita com o motor de busca, considerasse

[9] http://code.google.com/apis/ajaxsearch/documentation/

[10] http://www.json.org/java/

[11] PubMed, http://www.ncbi.nlm.nih.gov/pubmed

que esta seqüência corresponde ao nome de uma pessoa, que pode vir a ser o autor do conteúdo.

Desta maneira, são eliminadas seqüências como *"Save Search", "Contact Us",* e outras, que foram identificados por se enquadrarem na expressão regular. A Tabela 2 apresenta o resultado deste experimento.

Tabela 2. Resultados do Segundo Experimento

Legenda	**Total**
Nome presentes	100
Nomes identificados pelo programa	96
Nomes corretamente identificados	89
Precisão – *Precision*	93%
Abrangência – *Recall*	89%

Em relação aos resultados obtidos no primeiro experimento (Tabela 1) a precisão aumentou de 62% para 93%. A abrangência caiu de 97% para 89%. Se for utilizada *F-Measure,* para avaliar conjuntamente a precisão e abrangência, o resultado no primeiro experimento (Tabela 1) será de 75% e para o segundo experimento (Tabela 2) será de 91%, o que indica claramente uma melhoria nos resultados.

Foi ainda realizado um último experimento que consistiu em realizar a extração de nomes de pessoas de páginas (em Inglês e Português) retornadas a partir de consultas feitas no *Google*. Os experimentos foram realizados em páginas que contém os nomes dos autores que estavam entre os primeiros resultados apresentados pelo motor de busca*,* para uma consulta contendo a expressão "*Alzheimer*".

Uma vez identificadas estas páginas, foi realizado o mesmo procedimento de extração do primeiro experimento (passo 1 - seção 3.3), seguido do processo realizado no segundo experimento (passo 2 – seção 3.3). O resultado deste terceiro experimento é mostrado na Tabela 3. Cabe ressaltar que estas páginas apresentavam um menor grau de estruturação muito menor em relação aquelas dos experimentos anteriores, assim, já era esperado que os resultados fossem piores. Pode se notar que a pesquisa na *Web* com o uso dos padrões mostrados em (2) colaboram para o aumento da precisão (de 1% para 24%).

Tabela 3. Resultados do Terceiro Experimento

	Sem padrões	**Com padrões**
Nomes Presentes	19	19
Nomes identificados pelo programa	200	79
Nomes corretamente identificados	19	19
Precisão – *Precision*	1%	24%
Abrangência – *Recall*	100%	100%

5 Considerações Finais

Considerando que autoria é um importante indicador de qualidade de uma página *Web*, o artigo descreve um conjunto de técnicas procuram auxiliar na identificação dos nomes dos autores do conteúdo em páginas *Web*. O resultado deste processo de extração é incorporado a uma ontologia onde estão definidos critérios de qualidade para páginas *Web*.

A extração é feita desconsiderando aspectos relacionados à estrutura de uma página, o que é menos freqüente em trabalhos da área de extração de informações. Além disto, o conteúdo da *Web* é utilizado como forma de validar e qualificar a extração realizada. Os resultados, embora preliminares, demonstram que a estratégia adotada, pode conduzir a bons resultados. É preciso considerar que os requisitos expressos na seção 3.2 dificultam o processo de extração.

Neste sentido, constatou-se que é preciso buscar formas de identificar as expressões (“é autor de”, por exemplo) usadas neste processo, já que no trabalho, estas expressões foram identificadas a partir de outros trabalhos e durante a realização de alguns experimentos prévios.

Neste sentido, é necessário aprimorar as *queries* utilizadas para realizar consultas na *Web*. Por exemplo, uma *querie* "*papers by NE*" identifica mais claramente um autor do que uma *query* como "Mr. NE". O uso de operadores booleanos, pode auxiliar na produção de resultados melhores. Por exemplo, "*papers by NE*" AND "*Mr. NE*" podem garantir que está se identificando o nome de uma pessoa, e não o de uma companhia, como em *Papers by The X Paper Company*. Assim, a geração das *queries* com as expressões e seus conectores booleanos são importantes, também na construção e posterior utilização da *Web* como dicionário.

Ainda em trabalhos futuros, será preciso determinar, de forma mais precisa, dentre os possíveis autores que são hoje identificados, aqueles que são de fato os autores. Neste sentido a proposta é aplicar alguma das heurísticas usadas em [5]. Complementarmente, observando também os resultados e conclusões apresentados em [5], é necessário identificar

expressões, que estejam relacionadas aos autores da página (padrões relacionados à presença de dados bibliográficos, a identificação do nome/perfil do autor de um blog ou página pessoal).

Outro trabalho futuro consiste em tentar validar a autoria. Embora isto não seja possível de ser feito com total precisão, no cenário da *Web* atual, pode-se pensar em pelo menos indicar aos usuários a presença ou ausência de informações na *Web* que auxiliem na comprovação da autoria, de forma a eliminar possíveis *HOAX* cuja autoria é atribuída a determinadas pessoas. Para tanto, a idéia é buscar por referências ao autor em outros lugares da *Web*.

6 Agradecimentos

Este trabalho é parcialmente financiado pelo CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico/Cyted SALUS e CNPq Pro-Sul AVAL-SAÚDE.

7 Referências

[1] Covella, G. J. and Olsina, L. A. 2006. Assessing quality in use in a consistent way. In *Proceedings of the 6th international Conference on Web Engineering* (Palo Alto, California, USA, July 11 - 14, 2006). ICWE '06, vol. 263. ACM, New York, NY, 1-8. DOI= http://doi.acm.org/10.1145/1145581.1145583

[2] Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. 2005. Unsupervised named-entity extraction from the Web: An experimental study. *Artif. Intell.* 165, 1 (Jun. 2005), 91-134. DOI= http://dx.doi.org/10.1016/j.artint.2005.03.001

[3] Hasegawa, T., Sekine, S., and Grishman, R. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association For Computational Linguistics* (Barcelona, Spain, July 21 - 26, 2004). DOI= http://dx.doi.org/10.3115/1218955.1219008

[4] Karkaletsis, V., Karampiperis, P., Stamatakis, K., Labský, M., Ruring; žička, M., Svátek, V., Cabrera, E. A., Pöllä, M., Mayer, M. A., Leis, A., and Villarroel Gonzales, D. 2008. Automating Accreditation of Medical Web Content. In *Proceeding of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial intelligence* M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. Avouris, Eds. Frontiers in Artificial Intelligence and Applications, vol. 178. IOS Press, Amsterdam, The Netherlands, 688-692.

[5] Kato, Y., Kawahara, D., Inui, K., Kurohashi, S., and Shibata, T. 2008. Extracting the author of web pages. In *Proceeding of the 2nd ACM* WICOW '08. ACM, New York, NY, 35-42. DOI= http://doi.acm.org/10.1145/1458527.1458537

[6] Laender, A. H., Ribeiro-Neto, B. A., da Silva, A. S., and Teixeira, J. S. 2002. A brief survey of web data extraction tools. *SIGMOD Rec.* 31, 2 (Jun. 2002), 84-93.DOI= http://doi.acm.org/10.1145/565117.565137

[7] Lichtnow, D., Pernas, A. M., Manica, E., Kalil, F., Palazzo, O. J. M. de; Leithardt, V. R. Q. 2010. Automatic Collection of Authorship Information for Web Publications. In: *Proceedings of 6th International Conference on Web Information Systems and Technolo*. (Valencia, Spain, April 7 - 10, 2010). v. 1. p. 339-344.

[8] Liu, Y., Chen, R., Yang, H. Web Information Extraction Based on Hierarchical Model. 2009. In: Internation Conference on Computational Intelligence and Software Engineering. p.1-5.

[9] Matuszek, C., Witbrock, M., Kahlert, R. C., Cabral, J., Schneider, D., Shah, P., and Lenat, D. 2005. Searching for common sense: populating Cyc™ from the web. In *Proceedings of the 20th National Conference on Artificial intelligence - Volume 3* (Pittsburgh, Pennsylvania, July 09 - 13, 2005). A. Cohn, Ed. Aaai Conference On Artificial Intelligence. AAAI Press, 1430-1435.

[10] Patrick, J., Li, Mi. Intelligent Clinical Notes System: An Information Retrieval and Information Extraction System for Clinical Notes. 2009. In: 11th International Conference in e-Health Networking, Applications and Services. p.108-115.

[11] Sarawagi, S. 2008. Information Extraction. *Found. Trends databases* 1, 3 (Mar. 2008), 261-377. DOI= http://dx.doi.org/10.1561/1900000003

[12] Sari, Y., Hassan, M. F., Zamin, N. 2009. A Hybrid Approach to Semi-Supervised Named Entity Recognition in Health, Safety and Environment Reports. In: International Conference on Future Computer and Communication. p.599-602.

[13] Silva, O. P. da & Souto, M. A. M. 2006 Implementação e Avaliação de Mecanismos de Coleta de Dados na Web. Relatório Final de Bolsista IC – Projeto PERXML UFRGS.

[14] Zhu, J. 2009. An Adaptive Approach for Web Scale Named Entity Recognition. In: *1st IEEE Symposium on Web Society.* (Lanzhou, China, Aug. 23-24, 2009). p. 41-16