A Navigational and Structural Approach for Extracting Contents from Web Portals

Débora A. Corrêa[1], Ana Maria de C. Moura[2], Maria Claudia Cavalcanti[1]

[1]Department of Computer Engineering
Military Institute of Engineering (IME) Praça General Tibúrcio 80
Praia Vermelha, Urca, Rio de Janeiro, RJ, Brazil

[2]Extreme Data Lab (DEXL Lab)
National Laboratory of Scientific Computing (LNCC)
Petrópolis, RJ, Brazil

deboradac@gmail.com,yoko@ime.eb.br, anamoura@lncc.br

Abstract. *In a semantic Web portal, contents are described and organized based on domain ontologies, and are usually extracted from traditional portals. However, with the increasing amount of information generated each day on the Web, updating semantic portals still represents a major challenge, since this task lacks mechanisms to extract and integrate information dynamically. This paper proposes a strategy to help promoting the interoperability between portals. It consists on the extraction of contents from different Web sites on a specific domain, aiming at the instantiation of a domain ontology, and then use it to update and/or populate a semantic portal. This is carried out through the analysis of the navigational and structural characteristics of traditional portals endowed with some semantic potentiality. In order to evaluate this strategy, a tool named NECOW was implemented. NECOW performance was compared to the Google advanced search mode, and showed promising results.*

1. Introduction

Due to the explosive growth, popularity and heterogeneity of the Web, current traditional portals have difficulties to deal with the maintenance of their pages. They are still very limited for exchanging, reusing and integrating contents of other portals, as well as they rarely present efficient information extraction strategies and metadata maintenance. More recently, many efforts have been devoted in the area of information extraction (IE), whose main goal is to produce structured data from Web pages, so that they become ready for post-processing.

Semantic Portals (SP) arose as an evolution of traditional portals [Brickley et al. 2002][Lausen et al. 2005] [Mäkelä et al. 2004], and emerged as an attempt to provide an informational infrastructure with semantic meaning. They are characterized by the use of ontologies, with the aim of providing more semantic expressiveness to their informational contents. This is achieved by the improvement of some tasks performed over their contents such as search, organization and classification, sharing, publishing and inference. Hence, besides using the same technologies usually used in the construction of traditional portals, they additionally use ontological languages (RDF[1] and OWL[2]) to better organize structure and provide information semantic meaning in the portal pages [Reynolds et al. 2004].

Despite the advantages of SP, additional techniques are required to ensure these portals can be automatically populated and updated, since many of them still depend on manual update mechanisms. Automatically updating a SP with contents from other traditional portals (sites) depends strongly on Web information extraction (IE) techniques. Due to the heterogeneity and lack of structure of Web traditional portals, access to this huge collection of information is still a challenge, and has been limited to browsing and searching.

Consider, for example, a SP on the education domain that provides information about academic institutions and their courses. When a student wants to collect information about courses from different institutions in Rio de Janeiro, such as UFRJ[3] and IME[4], usually she/he has to navigate through their respective portals. In order to have access to the UFRJ courses, it is necessary to navigate through a list of Web pages, structured in a completely different way from that of IME portal. In fact, this scenario illustrates how difficult it is to extract information from such portals, and consequently, how hard it is to exchange information among them and maintain an up to date semantic portal.

In the literature, some works have been developed in this direction. Makella et al. (2004) use the idea of multi-facets to improve search mechanism in SPs, supported by ontological reasoning capabilities. In [Lachtim et al. 2009], a light ontology on the educational domain is used as the basis for integrating information, developing and populating semantic portals. Although these works aim at enriching portals, and at providing contents with more semantic meaning, they do not contemplate automatic information capturing from other traditional portals (or sites) available on the open Web. In the latter work, an architecture was proposed to retrieve information from semantic Web sites based on domain ontologies, which is then used to integrate contents collected from different SPs. In the present work we extend this idea, since the focus here is on extracting information from traditional portals on a specific domain. This information is transformed into structured data and used to instantiate a domain ontology, which serves as the main basis to automatically instantiate a SP on a specific domain, contributing to its maintenance [Corrêa 2012].

This paper proposes a strategy to deal with the interoperability between portals, also considering the possibility to automatically instantiate a SP. This strategy is based on the instances found along the navigational and structural analysis of Web portals. In order to achieve these goals, we assume that the portals we are going to deal with have some semantic potentiality. This term is used here to refer to traditional portals, whose contents are organized according to a hierarchical structure, helping users to navigate through the subject categories of their interest. These portals, claimed to be potentially semantic, use a somewhat controlled vocabulary, and terms typically appear as links and menu items throughout the portal. Examples of such portals are DMOZ[5], Wikipédia[6], and IME[7]. The

[1] http:// www.w3.org/RDF/.

[2] http://www.w3.org/TR/owl-features/

[3] Universidade Federal do Rio de Janeiro - www.ufrj.br.

[4] Instituto Militar de Engenharia - www.ime.eb.br.

[5] http://www.dmoz.org/

[6] http://pt.wikipedia.org/

[7] http://www.ime.eb.br/

main contributions of this paper are: (a) the specification of a navigational strategy to facilitate the identification of new instances to feed a SP; and (b) the evaluation of the proposed strategy. To the best of our knowledge, it is the first work in the ontology-based IE field that follows a navigational strategy.

The remainder of this paper is structured as follows. Section 2 gives a brief description of some essential concepts that are used throughout the paper. Section 3 presents some related work. Section 4 describes our navigational strategy for automatically extracting contents from portals and populating an ontology, with a brief description of its main functionalities. Section 5 presents NECOW, an extraction tool that has been developed according to the strategy proposed, with an example to demonstrate its usage. Section 6 is dedicated to the tool evaluation, and finally section 7 concludes the paper with suggestion for future work.

2 Extracting Information from Web Portals with Semantic Potentiality

Some traditional portals do organize their contents according to a hierarchical structure, helping users to navigate through the subject categories of their interest. However, in this work, we develop a navigational strategy to extract information based on structures found on portals that present some *semantic potentiality*. We define such a portal as the one that contains one of the following characteristics: (i) has links, lists or tables and benefits from any kind of organization and hierarchy in its structure; and/or (ii) some of its pages are presented as a taxonomy, although not all of them.

While DMOZ and some academic portals such as those of IME and UFRJ are classified in this category, others such as DBLife[8], DBPedia[9], FreeBase[10] are considered more comprehensive collaborative portals, since they provide a wide set of services that help dissemination and sharing information.

Semantic portals make use of semantic Web technologies to improve important functionalities in a portal, such as search and organization. Among these technologies, ontologies are considered as the most significant ones, since they enable common understanding and sharing of a domain between humans, agents and applications. Ontologies are also crucial to organize SPs, grouping sites and documents in pre-defined sets, according to their contents.

Due to the great heterogeneity of structures embedded in Web pages, extracting relevant data from them is still a challenge. IE is a classic text mining technique, whose goal is to find some specific information in texts, by identifying information contained in non-structured information source. This information should be in agreement to a predefined semantics, so that it could be later stored and/or manipulated by several other sources.

In the literature, three important IE techniques are identified [Silva A.S. 2012]: i) wrappers; ii) those based on Natural Language Processing (NLP); and iii) those based on the Deep Web (DW). The first one aims at extracting information from structured or semi-structured data (such as HTML). They are based on their format, delimiters, typography and frequency of words. NLP aims at extracting information directly from unstructured texts, and depends on the natural language pre-processing such as in Ondux [Cortez et al. 2010] and JUDIE [Cortez et al. 2011]. Finally, those based on the DW aim at extracting

[8] http://dblife.cs.wisc.edu

[9] http://dbpedia.org

[10] http://www.freebase.com

information from forms and/or hidden tables that are not visible to the user, as in DeepPeep [Barbosa and Freire 2005] and DeepBot [Álvarez et al. 2007].

Wilmalasuriya and Dou (2010) wrote an interesting overview about ontology-based IE technologies, also exploring some related tools. However, to the best of our knowledge, none of the discussed works used a Web navigational strategy, nor focused on the maintenance of semantic portals.

3. Related Work

A challenging research topic for the Web researcher's community is the interoperability between portals and their automatic instantiation. The literature points out to some works that use semantic Web technologies to exploit this topic, although in different contexts.

Lachtim et al. (2009a, 2009b) created an educational semantic portal, which is populated and integrated with contents extracted from Web semantic pages within the same domain. In [Suominen et al.2009] metadata and documents are obtained from contents published in Content Management Systems or from those manually annotated by the metadata editor SAHA [Kurki and Hyvönen 2010]. Later these metadata are submitted to an ontology to be validated and published in a semantic portal. The portal presented in [Hyvönen et al. 2009] creates its contents by making use of a set of metadata schemas and some specific tools. This population process enables producing and extracting contents from museums, libraries, files and other organizations, besides getting information from citizens as individuals and from national and international Web sources.

When compared with these works, our great differential consists on the semantic portal update with contents hosted in sites and/or Web portals with some semantic potentiality, and considering only their presentation and navigational structure, such as links, lists and tables. Hence, the update task in these portals allows these pages to be transformed from simple user pages into ones that are able to integrate and instantiate contents based on domain ontologies.

4. An Approach for Navigating and Extracting Information

This work extends the architecture proposed by Lachtim et al. (2009a). The latter aimed at creating a semantic portal, integrating and instantiating a domain ontology that supported a SP with contents extracted from Web semantic pages within the same domain. However, that architecture did not consider contents extracted from traditional portals or sites in the open Web. This work proposes a strategy to fill in this gap, as described along this section.

Figure 1 presents an overview of the proposed strategy. Mainly, the idea is to navigate through a list of sites with some semantic potentiality, on a specific domain. The navigation is guided by a subset (cropping) of a domain ontology (OB), which is represented in OWL. For each site in the list, *useful*[11] information is extracted to enrich the ontology, i.e., new potential instances of the OB classes, as well as new potential relationships between them (instances of OB object properties), are identified. A user validation of such new information is needed in order to remove eventual false positives. All this information is then transformed into RDF triples, which compose a new version of the OB ontology, here called OB′. The OB′ ontology may be used as input for the alignment with the already existing information in the current semantic portal. The main component of

[11] In the context of this paper, *useful* means all kind of information that is pertinent with the current domain.

this architecture concerns the IE, which gives the required support for populating portals. This component, as illustrated in Figure 1, is composed of modules, whose characteristics and functionalities are described next.

i.OB cropping: this step is responsible for loading a list of classes, instances and properties of the OB domain ontology, which will be the basis for the search of the *useful* instances of each page visited in a portal with semantic potentiality. The relationships between classes of the OB are also considered, since they guide the navigation along the portal pages. The navigation always starts from the most general class, defined by the user, and proceeds to the more specific ones. Additionally, the real name of each class, its label and its equivalent classes are very important for the navigation between the pages of the portal (see step iii). The instances of each class, as well as their equivalent instances (defined by the property *same as*) are also considered;

ii.Pre-categorization and identification of the initial page: a pre-categorization based on the title will be performed to limit the navigation defined by the step (i). If the initial page contains in the title a name similar to an instance of any OB class, the navigation will start from the next class of the OB. If this situation does not occur, the navigation will start from the first OB class;

iii. Navigation: this module is responsible for the navigation through the pages of each site previously defined by the user (stored in a configuration file). Its main goal is to search for links, within table or menu lists, through which OB' classes can be identified and corresponding new instances can be retrieved. It is composed of four sub-modules:

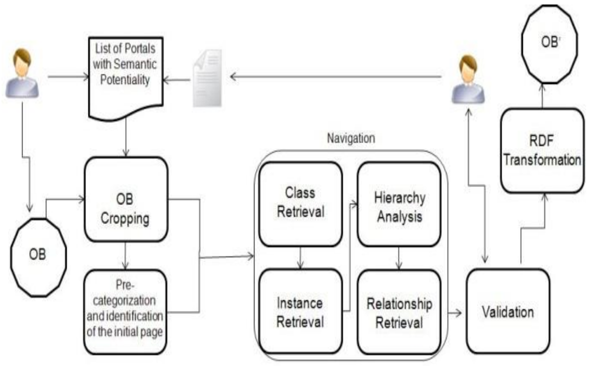


Figure 1. Navigation and Information Extraction Module.

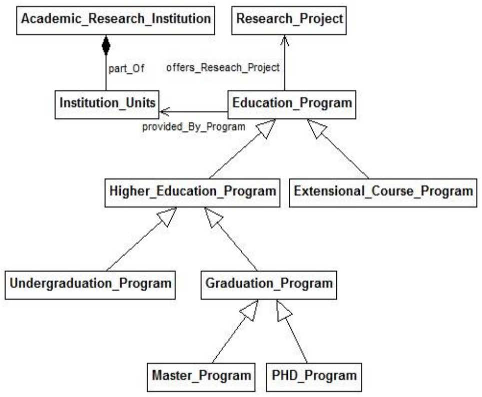


Figure 2. Fragment of the OBEDU ontology (OB)

A.Class retrieval: once the navigation starts, the system will search for links and labels that are similar to the desired OB' classes defined in step (i). These links will be considered as priority for navigation. Whenever a similar link is identified, the system verifies if it has already been visited. In the affirmative case, it will go on through the next link; otherwise, the link will be visited and its instances will be retrieved as described in the next step B;

B.Instance retrieval: for each OB' class similar link identified in step A, the corresponding target page is traversed in order to identify potential instances to that class. These instances should appear between *tags*, denoting links, lists and/or table items. Additionally, their label should have some similarity with the existing instances of the corresponding OB' class. During the navigation, all the information extracted is saved for later validation by the user (step iv);

C.Hierarchy analysis: in order to avoid duplicated instantiation in the OB´, hierarchies should be verified. This duplication typically occurs, for example, with a class and its subclasses. As an instance can instantiate a class and also its superclasses, the most specific one is chosen;

D.Relationship retrieval: associations between instances should be in accordance with the existing OB relationships. Hence, for example, in the ontology shown in Figure 2, the instances of "*Education_Program*" are associated with those of "*Academic_Research_Institution*" through the property "*provided_By_Program*". Among the new set of instances, such new associations are also identified, and later transformed into RDF triples (step v);

iv. Validation: this module is responsible for allowing the user to validate all the information extracted by the system during navigation. Even that one that may be considered as invalid is also saved, in order to be used later in a pre-validation process. This information can be confronted with the one that is retrieved later, during a posterior navigation;

v. RDF transformation: this module converts valid information into RDF triples, which will be included in the new ontology, the OB´. Actually, this corresponds to an empty crop of the OB, which is updated with the new instances extracted during navigation. OB´ triples can be submitted to an ontology alignment process with the OB ontology, and its instances will then be used to populate a semantic portal having the OB as its domain ontology. This alignment step is not in the scope of this paper.

5. NECOW: a Prototype Tool

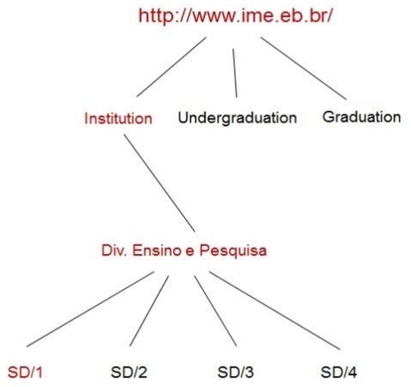
This section describes the prototype tool, named NECOW (Navigation and Extraction of COntents on the Web), developed with the objective to evaluate and test the strategy proposed in this work. It is a Web friendly tool developed in Java 1.6, and supported by some libraries (Jena[12], Jericho parser HTML[13], etc.). Navigation in NECOW starts from a portal Web link defined by the user, with the support of the base ontology (OB), which is loaded in memory and will help during all the navigation process for the search of classes and instances. It is worth observing that the strategy presented in section 4 is a generic proposal, and may be applied to other domains, for which there is a domain ontology. However, in order to show how this strategy is performed using NECOW, we will use an example in the educational domain, which is supported by the OBEDU ontology [Lachtim et al.2009a], which provides English and Portuguese vocabulary. A fragment of this ontology is presented in Figure 2. We also start our use case example with the IME institution, described through its portal, as shown in Figure 6.

When navigation starts through this portal, the html page source code of each page visited along the process is analyzed to verify if the label corresponding to the tags *title, link, item list* and *HTML tables* (*<title>,<a>, <li>* and *<td>*, respectively) has any similarity with a class and/or instance of the OB ontology. When this occurs, the corresponding tag labels are stored in a list. Hence, this navigation follows the same principle of a crawler, where only links associated with the OB are used in the process. Additionally, each candidate instance selected is stored in a list associated with its class

[12] http://jena.sourceforge.net/ontology/ - used to manipulate ontologies.

[13] http://jericho.htmlparser.net/docs/index.html -used to extract information from Web pages during navigation.

(Figure 3). However, for the tag<*a*> the process is different: the *href* (attribute that contains an *url*) and the tag *label* are extracted. For the labels of the tag<*a*> the same analysis for the labels is done, while for the *href* content, the links containing a label or a word (in its own kink) that has any OB class are extracted and stored in a list of links. Later these links are structured dynamically as an *n-ary* tree to identify the navigation path and the relation between these links.



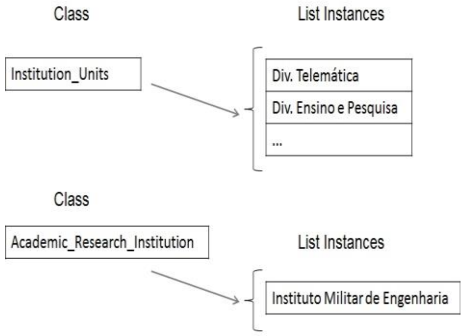


Figure 3. Instances of each class   Figure 4. A Tree representing the portal links

This tree also indicates which instances have been extracted for that link and their original classes. Figure 4 presents the navigational structure of the IME portal, whose nodes correspond to the links found and that will be visited during the NECOW execution.

In Figure 6, the first link is followed by the *href content* *http://www.ime.eb.br*. When accessing the page referenced by this link, we find "Instituto" (Institution, in English) as the content of the *href* *http://www.ime.eb.br/index.php?option=com_content&view=article&id=219&Itemid=3*. Performing again the previous step, the link "Instituto" directs us to the link "Div. Ensino e Pesquisa, SD1" (SD1 is the name of a Teaching and research division at IME), etc. (Figure 5). Navigation is performed in a *breadth-first* search, and before storing a link the system verifies the category of the link: if it is identified as a relative link [14] it is concatenated according to its domain, otherwise the absolute[15] ones are stored integrally. The similarity degree between words is calculated according to the edition cost, using the Levenshtein algorithm [Navarro 2001]. The edition cost consists on obtaining the number of operations (insertion, delete or modification) required to transform a word into another, one character at a time. This cost comprehends an interval between 0 and the size of the biggest word. Zero indicates the words are the same, and the larger the value, the greater the number of operations performed, and consequently, more different the words are.

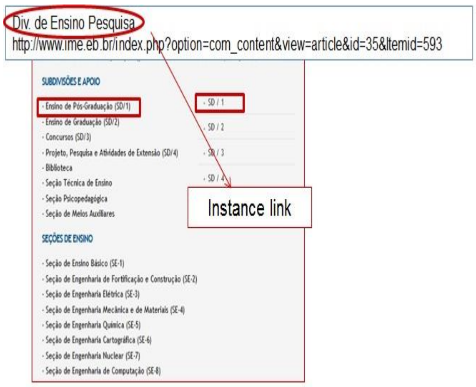
During navigation, the associations found between the links and their respective lists and tables (HTML) are compared with those of the OB, as illustrated in Figure 7. Those that are in accordance are stored, and at the end of the navigation they are presented to the users as RDF triples to be validated. When the associations are not identified during navigation, the relationships based on the ontology associations are suggested to the user.

[14] Its address is written in a summary way, containing only their directory names.

[15] The address is written integrally.

6. Evaluation and Results

This section is dedicated to the evaluation of the NECOW tool. Although the proposed approach is generic and can be applied to any specific domain, our test scenario has been developed on the educational domain, according to the ontology partially described (OBEDU) (Figure 2). The tests applied to NECOW aim at evaluating the tool efficiency with respect to the data extracted from the Web, in terms of precision and recall measurements. The extraction results were then analyzed and some were considered as valid instances to update the POSEDU portal [Lachtim et al. 2009b].


Div. de Ensino Pesquisa
http://www.ime.eb.br/index.php?option=com_content&view=article&id=35&Itemid=593

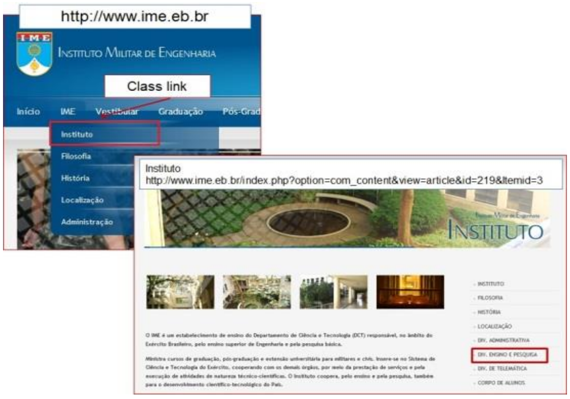
SUBDIVISÕES E APOIO

- Ensino de Pós-Graduação (SD/1)
- Ensino de Graduação (SD/2)
- Concursos (SD/3)
- Projeto, Pesquisa e Atividades de Extensão (SD/4)
- Biblioteca
- Seção Técnica de Ensino
- Seção Psicopedagógica
- Seção de Meios Auxiliares

SEÇÕES DE ENSINO

- Seção de Ensino Básico (SE-1)
- Seção de Engenharia de Fortificação e Construção (SE-2)
- Seção de Engenharia Elétrica (SE-3)
- Seção de Engenharia Mecânica e de Materiais (SE-4)
- Seção de Engenharia Química (SE-5)
- Seção de Engenharia Cartográfica (SE-6)
- Seção de Engenharia Nuclear (SE-7)
- Seção de Engenharia de Computação (SE-8)

Figure 5. Navegational structure of the IME portal



Instituto
http://www.ime.eb.br/index.php?option=com_content&view=article&id=219&Itemid=3

O IME é um estabelecimento de ensino do Departamento de Ciência e Tecnologia (DCT) responsável, no âmbito do Exército Brasileiro, pelo ensino superior de Engenharia e pela pesquisa básica.

Ministra cursos de graduação, pós-graduação e extensão universitária para militares e civis. Insere-se no Sistema de Ciência e Tecnologia do Exército, cooperando com os demais órgãos, por meio da prestação de serviços e pela execução de atividades de natureza técnico-científicas. O Instituto coopera, pelo ensino e pela pesquisa, também para o desenvolvimento científico-tecnológico do País.

Figure 6. IME portal and links to the tree of links

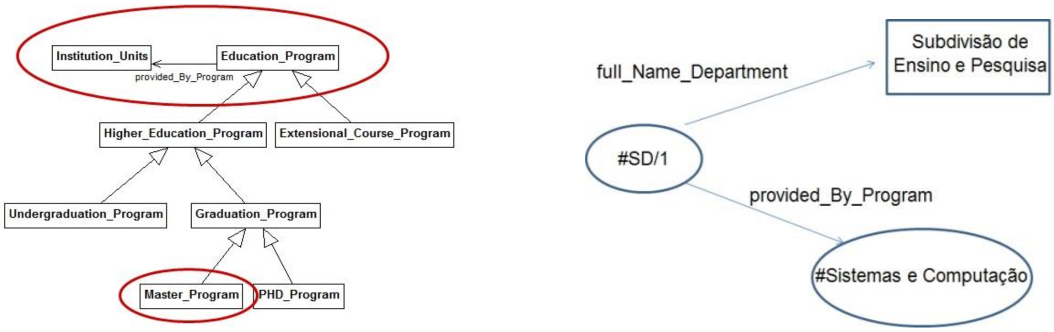


Figure 7. Instances found at IME portal related to the OB ontology

6.1. Test Scenario

At the time of developing this work we did not find in the literature a navigation tool with similar navigation strategy as NECOW, in order to use it as a comparison platform to evaluate our tool efficiency concerning its usefulness. Therefore, we evaluated NECOW navigation results by comparing them with those obtained from Google advanced search mode, over the same set of sites. In order to calculate the results precision and recall for each tool, we considered navigation and information extraction performed manually by a set of users, as the *gold standard reference*. Manual navigation was performed by a group of 20 users. Each user was told to select manually all the information considered relevant from a given list of portals. Both NECOW and Google also browsed the same list of portals. These portals were previously selected, taking into account that they all belong to a similar domain. Additionally, a related domain ontology (OB) was also chosen. It is worth observing that the automatic selection of portals (with semantic potentiality) on a specific domain is not in the scope of this paper. We assume these sites have been submitted to a previous analysis, to ensure they had some semantic potentiality, and that they could contribute with useful information for the OB ontology. Navigation was accomplished the same way in all procedures (manual, NECOW and Google), taking into account the same order in which classes and relationships were organized in the ontology.

The manual procedure was performed as follows. Initially we distributed a guide to the users containing a list of items to be considered, which includes: a list of links of educational portals to visit; a starting point page for each link in the visiting list; a list of the ontology terms they should search in each link; a list of similar terms to be used in an extended search expression. Additionally, while following such instructions, each user filled in a form with information concerning the ongoing manual navigation, such as: the page in which he/she was navigating; the terms (generic and /or specific) used to arrive at that page, as well as the path used; and also if these terms were linked to other pages. At the end, a set of new instances to the base ontology classes were documented by such users, and these were taken as the *gold standard reference*. With respect to the navigation procedures with tool support (Google and NECOW), the result set was compared to the *gold standard reference*. Navigation and information extraction executed by Google starts with the submission of a search expression built based on each ontology class (and its equivalent classes). Then, besides analyzing the returned page with a list of URLs, the user all the pages pointed by each URL are also analyzed. The search expression that was submitted to Google was manually built as follows: a class name followed by its equivalent classes separated by the OR operator, and followed by the (initial) link site. An example of one of the used expressions is: *Institution Unities OR Institutes OR Faculties - http:ime.eb.br.*

NECOW navigation and information extraction starts with a given web page (e.g. http://ime.eb.br), as explained in detail in section 5. Different from Google search mechanism, NECOW crawls through the site with the help of a chosen ontology structure (also a user input choice), and the user does not need to navigate through links. The candidate instances are suggested by NECOW at the end of its execution. Our tests have been performed over twenty specific sites on the educational domain, corresponding to some Brazilian universities. This number of sites was defined based on [Lachtim et al. 2009a] and [Navarro 2001]. Both works describe similar experiments concerning manual navigation, whose evaluation process is hard and tiresome, since it can generate a large number of instances. Based on preliminary tests, we considered as candidate instances (i.e., the ones that might be useful to be included in the portal) those that presented a similarity degree between 0 and 0.5, a value calculated according to the edition cost algorithm.

The experimental tests aimed at comparing NECOW results with the results obtained from Google, taking into account the *gold standard reference*, i.e., the results obtained with the manual navigation procedure. The set of results were the base to calculate the *precision* (P), *recall* (R) and *F-Measure* (F) coefficients, defined respectively by: $P = \frac{|Ra|}{|A|}$, R= $\frac{|Ra|}{|Ri|}$, $F = 2 * \frac{R*P}{R+P}$, where: Ra corresponds to the number of the relevant information instances retrieved by either NECOW or Google; A is the number of the information instances retrieved respectively by each tool; and Ri is the number of relevant information instances obtained with the *gold standard reference*.

6.2. Evaluation and Discussion

The portals were grouped according to the recall values, due to what was observed during preliminary tests: when we decreased the superior similarity threshold value, aiming at increasing precision, this one and the recall itself decreased considerably. Besides, many valid results were returned as invalid (false negatives). A more detailed analysis of the invalid results (true negatives at first) showed that some of them could be possibly evaluated as relevant if confirmed by the user. Since these *doubtful* results do not influence recall, but only precision, they have not been classified as invalid, in the invalid list.

The categories defined for recall have been defined as: (1) High recall (HR): results defined in the interval [0.70, 1]; Average recall (AR): results defined in the interval (0.40, 0.69]; Low recall (LR): results defined in the interval (0, 0.40]. Figure 8 shows the unified results obtained from Google and NECOW. Taking into account the *gold standard reference*, NECOW obtained the best recall results for most portals (UFJF, UFMG, PUC-Rio, UNESP, UFLA, UFF and UFG), whereas it presented the same results as Google for the other portals. A brief analysis of these portals, also considered as well *behaved portals* (HR), lead us to conclude that most of them present a good navigation and presentation structure, i.e., those that follow the basic HTML best practices. Furthermore, the terms used by them to describe the domain are quite close to those present in the OB. This explains the great difference observed with UNESP university portal, which reached 0,91 from NECOW, and 0,18 from Google. Additionally, we also observed the presence of many valid links and well defined labels in the portals, such as in IME, UFMG, UFRGS, UFC, UFJF and PUC-Rio universities. Considering the categories defined above and the lower recall obtained by both tools, we remarked that some portals have been classified differently by NECOW and Google. Similar situations occurred to UFLA (0.58) and UFG (0.46) that were classified as AR by NECOW, but as LR by Google (0.16 and 0.21, respectively); and to UFF classified as LR by NECOW (0.30) and as AR (0.47) by Google.

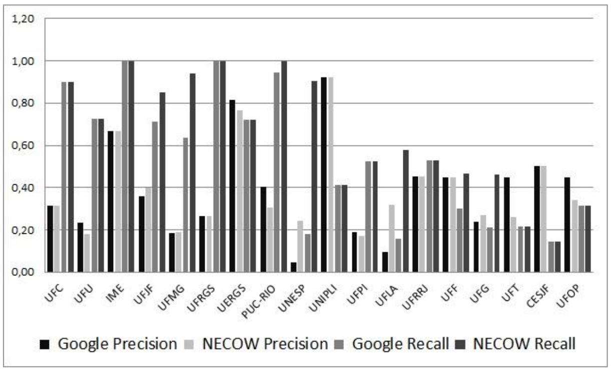


Figure 8. NECOW and Google results compared with the *gold standard reference*

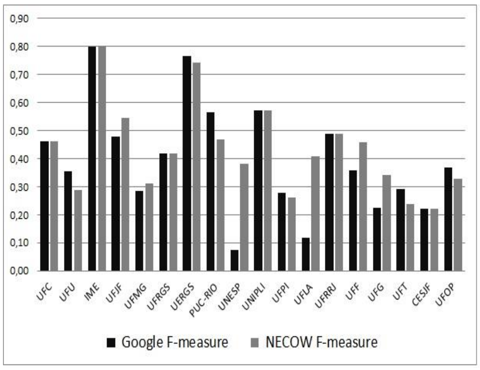
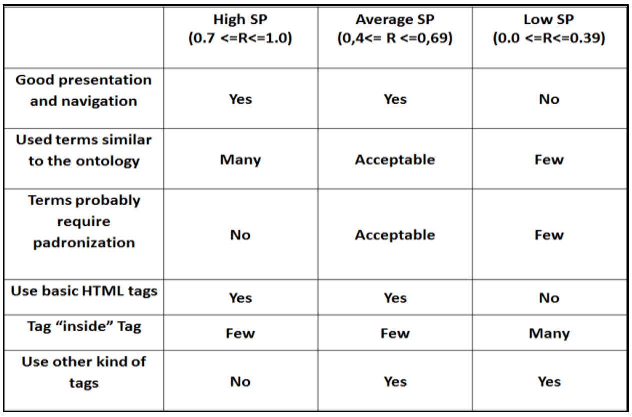


Figure 9. NECOW and Google *F_ measure* results

A detailed analysis of these values may justify such low results: presence of some portal internal links that redirects the server link, taking NECOW to invalid links; links that are not similar nor equivalent to the terms of the OB; terms that do not follow their standard usual connotation, or even links without any associated page; some information in the domain context were available through text indentation (TABs), difficult for the NECOW parser to find the desired information; use of <*link*> tag, instead of <*a*> tag, usually used for style sheets that, in conjunction with the use of frames, are not considered by NECOW yet. Two of our portal list fell in these cases. Precision results were lower than the recall ones. NECOW presented a discrete advantage in comparison to Google (UFJF, UNESP, UFLA and UFG), while for the others they presented similar results, except for UNIPLI, where we can observe a precision improvement (0.91). Actually, this portal presented a very few number of informational contents compatible with the OB concepts and hence, the information captured by NECOW was worse. Figure 9 presents the *F_measure* graphic. Its maximum values corresponded to the portals that had also the highest recall values (IME, UFRGS). From the analysis carried out along this work we concluded that NECOW obtained best recall results than Google. Additionally, it was possible to list some characteristics of portals with semantic potentiality, according to their classification into high, average and low categories, summarized in Table 1.

Table 1. Classification of Portals according to their Semantic Potentiality (SP).



	High SP (0.7 <=R<=1.0)	Average SP (0,4<= R <=0,69)	Low SP (0.0 <=R<=0.39)
Good presentation and navigation	Yes	Yes	No
Used terms similar to the ontology	Many	Acceptable	Few
Terms probably require padronization	No	Acceptable	Few
Use basic HTML tags	Yes	Yes	No
Tag "inside" Tag	Few	Few	Many
Use other kind of tags	No	Yes	Yes

7. Conclusion

This paper extends a previous work [Lachtim et al. 2009a], by proposing a strategy to collect contents from sites and/or traditional portals with some semantic potentiality within a specific domain, in order to instantiate an existing domain ontology that supports a semantic portal in this same domain. Actually, this strategy aims to facilitate the population and update procedures of semantic portals. In order to test and evaluate this proposal, a tool (NECOW) was implemented, and some tests were performed comparing it with Google advanced search tool, having as reference set the manual navigation performed by a group of users. It was possible to observe that manual navigation is usually more precise, and that the lack of structure in many portals design turns navigation and automatic extraction very hard. However, the good recall results obtained with NECOW were promising. It may be considered as an interesting and powerful tool to complement other IE techniques based on natural language processing, in the attempt of (semi) automatically populating semantic portals. As future work we intend to use machine learning techniques to improve information extraction process, as well as test other algorithms to calculate similarity between strings during this process.

Acknowledgements. This work has been partially supported by CAPES graduation scholarship and by CNPq through its Institutional Capacity Program (Proc. 382.489/09-8) and Productivity Research fellowship program (Proc. 309307/09-0).

References

Álvarez M., Raposo J., Pan A., Cacheda F., Bellas F., Carneiro, V. (2007). DeepBot: A Focused Crawler for Accessing Hidden Web Content. *University of La Coruña*.

Brickley D., Buswell S., Matthews B. M., Miller L., Reynolds D., Wilson M.D. 2002. Semantic Web Advanced Development for Europe (SWAD-Europe). In Proc. of the *1st Int. Semantic Web Conf. on The Semantic Web,* Sardinia, pages 409-413, 2002.

Barbosa L., Freire J. (2005). Searching for Databases. 18th *International Workshop on the Web and Databases* (WebDB 2005), Baltimore, Maryland.

Corrêa D.A. (2012). An Approach for Extracting Contents Based on Structural and Navigational Characteristics of Web Portals (in Portuguese). *Master thesis*, IME, Rio de Janeiro, Brazil, April.

Cortez, Silva A., Moura E. (2010). Ondux: On Demand Unsupervised Learning for Information Extraction. Proceedings of the ACM SIGMOD International Conference on Management of Data.

Cortez, Silva A., Moura E., Laender A. (2011). Joined Unsupervised Structure Discovery and Information Extraction. *Proc. of the ACM SIGMOD Int. Conf. on Mmt. of Data.*

Hyvönen E., Mäkelä E., Kauppinen T., Alm O.,et al. (2009). CultureSampo - Finnish Cultural Heritage Collections on the Semantic Web 2.0. *Proc. of the 1st Int. Symposium on Digital Humanities for Japanese Arts and Cultures (DH-JAC-2009),* Ritsumeikan Univ., Kyoto, Japan, March.

Kurki J., Hyvönen E. (2010). Collaborative Metadata Editor Integrated with Ontology Services and Faceted Portals. *Workshop on Ontology Repositories and Editors for the Semantic Web, the Extended Semantic Web Conference ESWC*, Heraklion, Greece, CEUR Workshop Proceedings.

Lachtim F. A., Moura A. M. C., Cavalcanti M. C. (2009a). Ontology Matching for Dynamic Publication into Semantic Portals. *Journal of Brazilian Computer Society* (JBCS), ISSN: 0104-6500, vol 15. pp 27- 43, Mar.

Lachtim F. A., Ferreira G.N., Gama R., Moura A. M. C., Cavalcanti M. C. (2009b). POSEDU: a Semantic Educational Portal. *IEEE Multidisciplinary Engineering Education Magazine,* vol 4, nº 3, pp. 65-75, ISSN:1558-7908.

Lausen H., Ding Y., Stollberg M., Fensel D., Hernandez R., Han S. (2005) Semantic Web Portals: State-of-the-Art Survey. *J. Knowledge Management*, V.9, N.5. May, pp. 40-49.

Mäkelä E., Hyvönen E., Saarela S., Viljanen K. (2004). Ontoviews - a Tool for Creating Semantic Web Portals. *Inte. Semantic Web Conference,* Hiroshima, pp. 797-811.

Navarro G. (2001). A guided tour to approximate string matching. Univ. of Chile. ACM Computing surveys, vol. 33, no. 1.

Reynolds D., Shabajee P., Cayzer S. 2004. Semantic Information Portals. *ACM*, NY, May.

Reynolds D., Wilson M.D. (2002). Semantic Web Advanced Development for Europe (SWAD-Europe). *In Proceedings of the 1st Int. Semantic Web Conf. on The Semantic Web*, Sardinia, pages 409-413, 2002.

Silva A.S. (2012). Methods and Techniques for Information Extraction by Text Segmentation. *Proc. of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management*, Ouro Preto, Brazil, June 27-30.

Suominen O., Hyvönen E., Viljanen K., Hukka E. (2009). *HealthFinland - a National Semantic Publishing Network and Portal for Health Information*, Finland.

Wimalasuriya D.D. and Dou D.(2010). Ontology-based Information Extraction: an Introduction and a Survey of Current Approaches. J.Inf. Sci. 36, 3, June.