Ontology-based Process for Recommending Health WebSites

Edelweis Rohrer[1], Regina Motz[1], Alicia Díaz[2]

[1]Instituto de Computación, Facultad de Ingeniería,Universidad de la República, Uruguay
{erohrer, rmotz }@fing.edu.uy
[2]LIFIA, Facultad de Informática,Universidad Nacional de La Plata, Argentina
alicia.diaz@lifia.info.unlp.edu.ar

Abstract. Website content quality is particularly relevant in the health domain. A common user needs to retrieve health information that is precise, reliable and relevant to his/her profile. Website recommendation systems are an aid to get high quality health-related web sites according to the user's needs. However, in practice, it is not always evident how to describe recommendation criteria for health website. The goal of this paper is to describe, by an ontology network, the criteria used by a health website recommendation process. This ontology network conceptualizes the different domains that are involved in the *Salus* Recommendation Project as a set of interrelated ontologies[1].

Keywords: web site, recommendation system, ontology network.

1 Introduction

The use of the web by common people as a repository of information, especially in the health area, increases drastically day by day. This is a very worrying reality because many of health web sites do not contain data according to user´s necessities. For example, a technical content may not be a good quality reference for a person which is not able to understand technical vocabulary. On the other hand, the alternative medicine products that are offered on the Internet, the lack of quality controls (editorial boards) at the stage of production of the web site and the "lack of context", lead to information does not necessarily have to be false to harm [1]. Furthermore, the fact that the web is a very dynamic medium, once a person has obtained misinformation, then, it is very unlikely to be reversed by health professionals. In this sense, a decentralized, ontology-based recommender system can automatically give an evaluation about the quality of the sources according to the consumer's needs.

Quality in websites is determined by several diverse factors, some of which are general, and therefore, considered for any type of sites and for any domain. Such features include, for example, navigational aspects, user interface aspects, legibility

[1] This work was partially funded by: the SALUS/CYTED and PROSUL projects which are sponsored by the CnPq, Brasil and the CyTED, Spain. It is also supported by the PAE 37279-PICT 02203 which is sponsored by the ANPCyT, Argentina.

(size of letter, colors, images), performance aspects (time it takes to access to the site content), the correct functioning of the site, its conformity with standards of language use or of accessibility like those described in normative such as the Web Content Accessibility Guidelines of the W3C[2]. Some quality models that take these features into consideration are WebQual [2] and WebQEM [3].

On the other hand, in this work we focus on the quality that arises of the information value that the site provides and its adequacy for the consumer's context. The consumer´s context contains two domains: the user profile and the query goal. The first is described by properties as gender, age, employment among others. The second relates to the purpose for which information is needed, it can be, for example, buy a drug, selecting a doctor or write a school paper.

Our approach is to consider, in an integrated way, the specific health domain of interest (i.e. diagnosis, treatment, etc.), the dimensions of quality factors, the user's context and the criteria to assure that some information is in accordance to the goal of "fitness for use" for a consumer [4]. With this aim, we specify a process driven by an ontology network that leads to give a recommendation of suitability of web contents to a particular user who makes a specific query. The ontology network describes how to set up the quality factors and the recommendation criteria considering the specific domain, properties of websites and consumer goals.

This paper is organized as follows. Section 2 describes issues about quality assurance and recommender systems. Section 3 presents the *Salus* ontology network. In section 4 we discuss the process for web sites recommendation based on the *Salus* ontology network. Finally, we discuss conclusions and future perspectives.

2 Background on Recommender Systems and Data Quality

Recommender systems could be defined as systems that produce individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options [9]. Within the broad range of existent works, we will mention some of the more recent ones.

The approach presented by Porcel et al. in [11] consists of a recommendation model based on vectors that represent the resource scope and users interests, and then to match them. There were distinguished four different classes of recommendation techniques: (1) *content-based systems*, based on the terms used about resources (2) *collaborative systems* that consider the user preferences, (3) *Demographic systems* that represent the different user profiles and (4) *knowledge-based systems*, based on inferences about resources that satisfy the users. These authors proposed a hybrid approach that combines content-based and collaborative techniques. In [10], Oufaida and Nouali present a multi view recommender system that includes collaborative, social and semantic views of the user's profile, related to a set of resources semantically annotated. Recently, in [12], it is presented the construction of a recommender system which is described as an iterative process; where at each iteration a model representing the preferential characteristics for the recommendation

[2]http://www.w3.org/WAI/GL/

is obtained. The system is an ontology-based recommendation process that produces recommendations by applying content-based, context-aware and collaborative criteria.

Unlike the mentioned works, our proposal of recommendation process is strongly based on the quality assessment of the web contents. However, there exist some common aspects with them. Considering the classification of recommendation techniques given by [11], our proposal also matches user profiles and resources, although we rather combine content-based (web content properties) and demographic (user profiles) approaches. Furthermore, some aspects faced by [12], as considering context issues (i.e. the query situation at the moment the user makes a query) and the exploitation of ontological structures that underlie the recommendation process, are also considered in our proposal. In the next section we present the quality assurance onto-logy and how it is related in the ontology network in order to specify a recommendation process.

There are, basically, two ways of defining data quality: the first one uses a scientific approach and defines data quality dimensions rigorously, classifying them as dimensions that are or are not intrinsic to an information system [4]. The second one is a pragmatic approach aimed at defining data quality in an operational fashion [5]. Wang et al. [4] identified four data quality dimensions: (1) intrinsic data quality; (2) contextual data quality, which defines the quality of the information within the context of the task; (3) data quality for data representation, which determines if the system presents the information in a concise, consistent, understandable way; (4) data quality regarding data access, which defines quality in terms of the role of the information system in the provision of the data.

Within each dimension it is possible to identify several factors, including: for *intrinsic data quality dimension*: believability, accuracy, objectivity and for *context dimension*: value-added, relevancy, timeliness, completeness, among others. The domain expert is the one who decides which of these factors are relevant for a specific domain and she/he is who defines the appropriate metrics to measure these factors. Regarding Believability, two definitions are introduced in [6]: *Believability*, which is *the extent to which data is regarded as true and credible* and *Reputation*, which is *the extent to which data is highly regarded in terms of its source or content*. About this factor in health domain, it is important to take into account the existence of sites with certified quality labels, such as HON (http://www.hon.ch/), WIS (http://www.portalesmedicos.com/) and WMA (http://wma.comb.es/). For the readability factor, in [7], it is introduced different readability metrics that were created for different domains and user profiles. It sets the following definition: *Readability is what makes some texts easier to read than others*. There are a lot of readability formulas created for different authors, like FOG[3] and SMOG[4] grade levels. Here also, the decision on which formula to use, FoG or SMOG, must be taken by a domain expert. The first step is to specify a formal model that represents the factors involved in the acquisition of the quality of web data as well as the different metrics that can be applied. The main intention of measuring the data quality is to provide a quantitative meaning of quality dimensions. Metrics are these quantitative or categorical representations of one or more attributes [15].

---

[3]FOG grade level = 0.4 (average sentence length + hard words)
[4]SMOG grade level = 3 + ?polysyllable count

Our approach to face this challenge is the design of an ontological model inspired in our previous work [8] on web data warehouse quality, by modelling a generic ontology for quality factors, independent from the specific domain and the different types of web data sources. It is easily tailored to different user domains and different types of web data through its connection in the proposed ontology-network. In the next section we present the quality ontology in the context of an ontology network.

3 Salus Ontology Network

The *Salus* ontology network helps to obtain a reading recommendation of health-related web contents for a particular user. Specifically, it conceptualizes the different knowledge domains that are involved in a recommendation system in a shape of an ontology network. An ontology network is a collection of ontologies related together through a variety of different relationships such as mapping, modularization, and versioning, among others [13]. Accordingly, a networked ontology is an ontology included in such a network, sharing relationships with other ontologies. Intuitively, this implies to define the ontologies' content, but also to define metadata information about the networked ontologies. Ontology metadata refers to the information which is attached to the ontology itself, not to its content. This ontology metadata would cover ontology provenance, purpose and the relations with other ontologies and semantic resources. They are critical because they describe an ontology network as a whole.

Salus ontology network conceptualizes the following domains: specific health, web site, quality assurance, user context and recommendation. Each domain is represented by one or more interrelated ontologies.

- *Health domain* ontologies conceptualize the health domain. The core ontology may be an already existing ontology like UMLS[5] which models for example the impact, treatment, risk factors, diagnostic, effects, and phases of a disease. This ontology can be refined in terms of a specific disease i.e Alzheimer, and thus can be modelled the concept "Alzheimer treatment". *Salus* ontology network is specific to the health field, but it could be adapted to other domains just by changing the health ontology by other domain ontology.
- *Web Site domain ontologies* conceptualize the domain of webpages and particularly describe the web resources that will be considered to participate in a quality assessment. The main concepts of this ontology are *web resource* and *web resource property*. *A web resource* is any resource which is identified by a URL; for instance, it can be instantiated as a webpage which has attached content. *Web resource property* models the properties that can be attached to a *web resource*. For instance, possible properties of *web resources* could be the “author”, the “amount of words”, etc. Among these properties there is a particular one, the “hasTopic” property that relates concepts (web resources) from the *Web Site* ontology with concepts in the *Specific Health* ontology. The “hasTopic” property describes what a *web resource* is talking about. These kind of properties should be retrieved through a specific information retrieval mechanism, as it will be detailed below.

[5] http://www.nlm.nih.gov/research/umls/

- *Quality Assurance domain ontologies* conceptualize metrics, quality assurance specifications and quality assessments. *Metrics* are formula defined base on *web resource properties*. A *quality assurance specification* describes the different *quality dimensions*; for instance readability, precision, believability, completeness, timeliness, etc. The *quality assurance specification* associates to each quality dimension the suitable metric calculus. A *quality assessment* models the assessment of a particular web resource (i.e. a web document) for a particular quality dimension through a specific metric. It also models the obtained quality level.
- *Context domain ontologies* describe user profiles and query resources. The user profile conceptualizes user properties such as user age range, role, academic level, health domain expertise. among others. The query resource represents the context of the query. The main concept of the query resource is the query goal.
- *Recommendation domain ontologies* describe the different criteria of recommendation for a particular context (user and query situation) and quality dimensions and the obtained recommendation level.

Salus networked ontologies are interrelated (see in the upper of figure 1) by three different relationships: *uses*, *extends* and *describes* relationships:

- The *uses* relationship relates two ontologies by the import primitive. For example, this relationship occurs between the *Web Site* ontology and the *Specific Domain* ontology because of a *webpage topic* can be any concept at the specific domain ontology. In the *Salus* ontology network, *webpage topics* could be treatment, diagnostic, etc. Thus, “Alzheimer Treatment” is a topic of “Alzheimer Webpage”.
- The *extends* relationship describes a more specific ontology which is the specialization of a more general one. The clearer example is the *Alzheimer* ontology, which is a specialization of the *Health* ontology. For example, at the *Health* ontology conceptualizes: diagnostic, treatment, risk factors, etc; then these concepts are specialized in the *Alzheimer* ontology.
- The *describes* relationship defines the relations between a model and its metamodel. For instance, the *Web Site* ontology is an instantiation of the Web Site Specification ontology. The later is a meta-ontology for the former. Webpages are typical concepts at the *Web Site* ontology and are modelled by the *webpage* class. This class is an instance of *Web Resource* class which is defined at the *Web Site Specification* ontology. Another example is the property “has Author” that is defined at the *Web Site* ontology as an instance of the *Web Resource Property* class, which is also defined at the *Web Site Specification* ontology.

On the bottom part of the Figure 1 is shown an example of the resulting knowledge base where the content of the “Alzheimer webpage” was assessed to be recommended to the user “Paul”. The content associated to the “Alzheimer webpage” has "Alzheimer Treatment" and "Alzheimer Diagnostic" as topics. In this example the recommendation assessment took into account the “Believability” quality dimension, assessed by “provenance” metric, which uses the “has Author” property of the webpage. The recommendation assessment also considers the fact the user Paul is a teenager and the goal of his query is "looksFor". Later, in the section 4 more detail about the networked ontologies will be given.

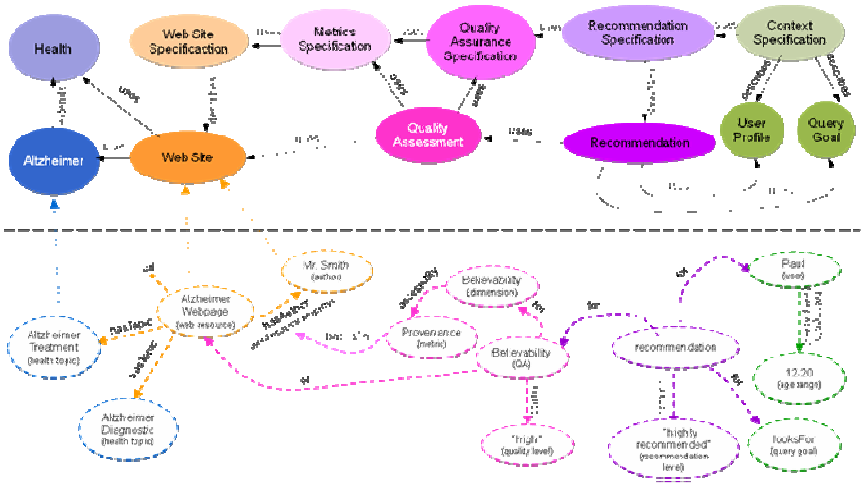

Fig. 1. An example of the *Salus* ontology network

4 *Salus* Recommendation Process

The *Salus* recommendation process covers the different tasks which have to be performed in order to recommend a set of web sites to a particular user. These tasks are organized in three different phases, namely, the start up of the recommendation system, the quality assessment of a set web pages and the execution of recommendation assessments. This process is characterized as an ontology-based process. Specifically, it is based-on the *Salus* ontology network described in previous section. During the execution of the *Salus* process, the *Salus* networked ontologies plays different roles: in some cases it helps to discovering knowledge domain units in the web pages (i.e. based on the specific health ontology), while in other cases, it helps to supporting quality or recommendation assessments. In the last cases, the *Salus* ontology network can be used to both: assist in the modelling and specification of a recommendation system and check the correctness of the resulting system specification. Particularly, this section will go in deep explaining the *Salus* ontology network during the recommendation start-up phase explanation.

4.1 Recommendation Start-Up Phase

The *recommendation system start-up phase* is in charge of preparing the information needed in order to recommend web pages. This phase consists of the tasks: web resource definition, quality criteria definition, recommendation criteria definition and context resource definition, schematized in Figure 2. Next, we will detailed them and show where, when and how the *Salus* ontology network is used.

Web Resource definition. It refers to the population of the Web site ontology according to a given set of webpages and their indexation based on the specific domain ontology. The *Web site* ontology is populated with *webpages* instances (one for each given webpages) and with properties that are involved in the newly defined instances; for example the "url" property is specified between a webpage and a URL. Then, these webpages are indexed according to the *Specific Domain* ontology; in *Salus*, it corresponds to the *Alzheimer* ontology. In this task the "hasTopic" property is specified between "Alzheimer webpage" and Alzheimer concepts, as it is shown in the figure 1. Then, in next task more properties will be discovered.

Quality criteria defnition. It refers to the definition of quality dimensions and metrics that will be supported by the recommender system. First of all, it have to be specified the repertoire of factors involved in the quality dimensions. Based on it, the definition consists on specifying which metric assets each quality factor and which are the possible obtained quality levels. Metrics are specified based on *web resource properties* (concepts of the Web site ontology). For example, when the "provenance" factor is instantiated, the "basedOn" property will be also instantiated in order to link the "provenance" factor with the "has Author" property. The "has Author" property has to be now specified as an instance of *web resource property* and it may be specified the metric used to capture this value. Then, the *Quality Specification* ontology has to be populated. Quality dimension concepts have to be instantiated. These quality dimensions are those supported by the recommender system. Each quality dimension concept at least has once defined the *assesedBy* property to link a quality factor to the metrics that enable its assessment. Quality dimension concepts also have defined the *assesTo* property to link a quality dimension to its possible quality levels. For instance, the dimension "Believability" has defined the *assesedBy* property which takes values in the "provenance" factor and the *assesTo* property to the set of strings: "high", "medium" and "low".

Recommendation criteria definition. It refers to the definition of recommendation criteria. Based on the quality criteria definition, a *recommendation definition* indicates which quality dimensions will be assessed and which *context resources* will be considered for a recommendation. *Context resources* are mainly *user properties* and *query resources*. The output of this task is a set of *recommendation rules* which specify the *recommendation level* for each assessed web page. These rules are like:

if recommendation definition(thisWebPage)
then thisWebPage **is** recommendationLevel(thisWebPage)

where *thisWebPage* is the currently processed webpage and the *recommendation definition(thisWebPage)* is described in terms of quality assurances and context. The recommendation level for a webpage is one of the scale values of the scales of recommendation levels of the recommender system (for example, "highly recommended", "strongly recommended"). Regarding, the example we have been followed along the paper, the rule below might be defined as follow:

if *BelievabilityQA(AlzheimerWebpage)* **assesTo** *"high"* **and** *Paul* **belongsTo** *12-20 age range* **and** *query goal* **is** *looksFor*
then *AlzheimerWebpage* **is** *highly recommended*

Context resource definition. It is in charge of defining those *context resources* that have to be taken into account to make a recommendation. These *context resources* will be identified in the recommendation criteria definition. Mainly, they are: the *user properties* and the *query resources*. The *user properties* are those that were already relieved at the recommendation criteria definition task and will be populated at the moment of registering a user at the recommender system: For instance, if at the recommendation criteria definition was specified the user property *belongs*, when she is registered to the system, this property is instanciated between *Paul* and *12-20 range*. The query resources refers to *query attributes* like *query goal*.

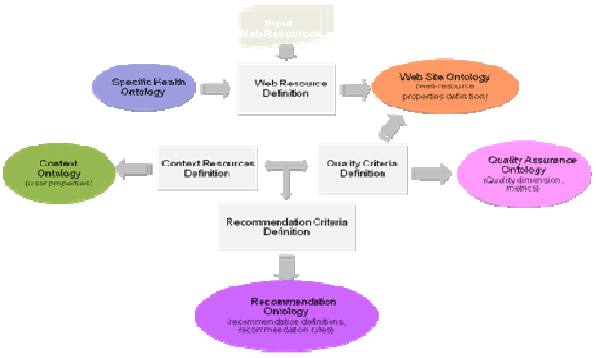

Fig. 2. Salus recommender start-up

4.2 Web Page Quality Assessment Phase

After the recommendation start-up phase, the quality assessment of a set of web resources can be done. First of all, the web resources will be pre-processed to determine their properties and populate the web site ontology. The metrics over factors involved in a quality definition, determine the values of the web resource to be considered in the criteria of recommendation for this dimension of quality. For example, in the definition of the dimension “Believability” is used the “provenance” factor which refers to the author of the webpage, i.e. the “Alzheimer webpage” should have associated the “hasAuthor” property. Therefore, it have to be determined which information retrieval process have to be performed in order to discover these new web resource properties. The retrieved information will be used to complete the population of the Web site ontology. Thus, the *hasAuthor* property can be defined between the “Alzheimer webpage” and “Mr. Smith”. In this phase, a set of specific domain web resources (webpages) will be assessed in order to determine their quality level. The quality assessment execution involves calculating the quality level of each web resource for each quality dimension. For that, the corresponding metric is executed and thus, it is determined the quality level of a web resource. In this phase, the quality assurance ontology is populated, mainly, by adding instances of the quality assessment and linking them with the web resource and the quality level. At this moment, the concept “BelievabilityQA ” is instantiated as an individual of the *Quality*

Assessment class and the *obtains* property is defined between “BelievabilityQA” and the "high" quality level.

4.3 Recommendation Assessment Phase

A user query is the trigger of this *recommendation assessment phase*. When a logged on user makes a query, the recommendation system evaluates the *recommendation rules* in order to determine the recommendation level. All those web resources, which assets to an appropriated level for the considered user, will be recommended.

The evaluation of the recommendation rules is based on the user profile, the quality level of the considered web resources and the query resources. Both, the user profile and the web resource quality level, have been calculated in the previous two phases. Query resources have to be discovered at this moment. The output of this phase is a set of recommended web resources to a particular user query. The figure 4 summarizes the recommendation assessment phase.

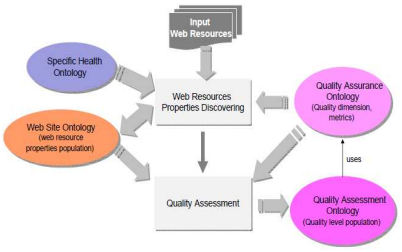

Fig. 3. Salus Quality Assessment

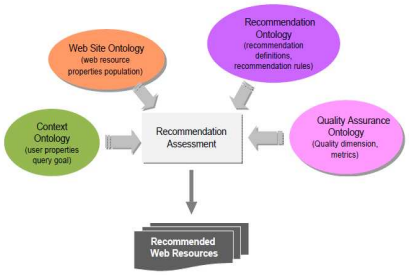

Fig. 4. Salus Recommendation Assessment

5 Conclusions and future work

In this paper we have introduced a novel approach which uses an ontology network to assist the modelling and execution of a website recommendation system. It is a quality-based approach to get the more adequate websites for a consumer and context.

We have described *Salus* ontology network that models the different domains related to a recommendation system. Moreover, we showed how this ontology network can be tailored, to specific health domains and user points of views. The main aim of this design was to obtain a flexible model that was not dependent on any particular mechanism of websites content evaluation, such as a specific quality metric or health domain. Whenever it is required to assess a different quality dimension or to consider another health domain, new extensions of web site, quality and recommendation ontologies might be added, keeping up the core model intact.

In addition, a valuable feature of driving the recommendation process by ontologies is the property of checking the consistency among concepts and relationships that allows one to detect inconsistencies at the design phase. Based on the intrinsic properties of ontologies, the model provides a high level abstraction that allows specifying in simple way relations between dimension and metrics for defining

quality assurance. Besides, it worth to mention that having an ontology-based recommendation system implemented using OWL language and SWRL rules, is helpful to validate the resulting configuration of the recommender system. These tools offers the possibility of defining restrictions and Horn-like rules that have to be hold in order to achieve consistent specifications of quality or recommendation assessments, detecting anomalous specifications.

Starting from the presented design, good practices on Ontology Engineering lead to evaluate the model in an interaction between ontology engineers and domain experts. From this evaluation, it is expected to obtain a feedback to reach a final refinement of the structures which compose the ontology network.

References

1. Gunther Eysenbach, T.L.D.: Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. BMJ 317,1496-1502,1998.
2. Barnes, S., Vidgen, R.: Webqual: An exploration of web-site quality. In: Proceedings of the Eighth European Conference on Information Systems, Vienna, July 3, 2000.
3. Mich, L., Franch, M., Inverardi, P.N., Marzani, P.: Choosing the "rightweight" model for web site quality evaluation. In Lovelle, J.M.C., Rodríguez, B.M.G., Aguilar, L.J., Gayo, J.E.L., del Puerto Paule Ruíz, M., eds.: ICWE, LNCS, Springer. Vol. 2722, 334-337, 2003.
4. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. 12(4) 5-33, 1996.
5. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Commun. ACM 39(11) 86-95,1996.
6. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM 45(4) 211_218, 2002.
7. Dubay, W.H.: The principles of readability. Costa Mesa, CA: Impact Information (2004)
8. Llambías, G., Motz, R., Toledo, F., de Uvarow, S.: Learning to get the value of quality from web data. In Meersman, R., Tari, Z., Herrero, P., eds.: OTM Workshops. , LNCS, Springer. Vol. 5333, 1018-1025, 2008.
9. Burke, R. Hybrid recommender systems. User Modeling and User-Adapted Interaction, 12(4), 331–370, 2002.
10. Oufaida, H. and Nouali O. Exploiting Semantic Web Technologies for Recommender Systems. A Multi View Recommendation Engine. Proc. of the ITWP'09, Pasadena, California, USA, July 11 – 17, 2009.
11. Porcel, C., Moreno, J. and Herrera-Viedma E. A multi-disciplinar recommender system to advice research resources in University Digital Libraries. Expert Systems with Applications. Volume 36, Issue 10, 12520-12528, 2009.
12. Bellogín, A., Cantador, I., Castells, P. and Ortigosa, A. Discerning Relevant Model Features in a Content-based Collaborative Recommender System. Preference Learning. Edited by Johannes Fürnkrans and Eyke Hüllermeier. Springer-Verlag., 2010.
13. Suárez-Figueroa, M.C., Dellschaft, K., Montiel-Ponsoda, E., Villazon-Terrazas, B., Yufei, Z., de Cea, G.A., García, A., Fernandez-Lopez, M., Gomez-Perez, A., Espinoza, M., Sabou, M.: Neon deliverable d5.4.1. Neon methodology for building contextualized ontology networks. Technical report, NeOn Project, 2008..
14. Olson J. Data Quality: the Accuracy Dimension. Morgan Kaufmann, 2003.
15. Angelica Caro, Coral Calero, Ismael Caballero, Mario Piattini: A proposal for a set of attributes relevant for Web portal data quality. Software Quality Journal 16(4): 513-542, 2008.