CORE: a tool for collaborative ontology reuse and evaluation

Conference or Workshop Item

CORE: A Tool for Collaborative Ontology Reuse and Evaluation

Miriam Fernández, Iván Cantador, Pablo Castells
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Campus de Cantoblanco, 28049 Madrid, Spain
{miriam.fernandez, ivan.cantador, pablo.castells}@uam.es

ABSTRACT
Ontology evaluation can be defined as assessing the quality and the adequacy of an ontology for being used in a specific context, for a specific goal. In this work, a tool for Collaborative Ontology Reuse and Evaluation (CORE) is presented. The system receives an informal description of a semantic domain and determines which ontologies, from an ontology repository, are the most appropriate to describe the given domain. For this task, the environment is divided into three main modules. The first component receives the problem description represented as a set of terms and allows the user to refine and enlarge it using WordNet. The second module applies multiple automatic criteria to evaluate the ontologies of the repository and determine which ones fit best the problem description. A ranked list of ontologies is returned for each criterion, and the lists are combined by means of rank fusion techniques that combine the selected criteria. A third component of the system uses manual user evaluations of the ontologies in order to incorporate a human, collaborative assessment of the quality of ontologies.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, retrieval models, selection process*.

General Terms
Algorithms, Measurement, Human Factors.

Keywords
Ontology evaluation, ontology reuse, rank fusion, collaborative filtering, WordNet.

1. INTRODUCTION
The Semantic Web is envisioned as a new flexible and structured Web that takes advantage of explicit semantic information, understandable by machines, and therefore classifiable and suitable for sharing and reuse in a more efficient, effective and satisfactory way. In this vision, ontologies are proposed as the backbone technology to supply the required explicit semantic information. Developing ontologies from scratch is a high-cost process that requires major engineering efforts, even for a medium-scale ontology. In order to properly face this problem, we believe efficient ontology reuse and evaluation techniques and methodologies are needed. The lack of appropriate support tools, and the lack of automatic measurement techniques for certain ontology features are often a barrier for the implementation of successful ontology reuse methods.

In this work, we present CORE, a Collaborative Ontology Reuse and Evaluation system. This tool provides automatic similarity measures for comparing a certain problem or Golden Standard to a set of available ontologies, retrieving not only those most similar to the domain described by the Golden Standard, but the best rated ones by prior ontology users, according to the selected criteria. For similarity assessment, a user of CORE selects a subset from a list of comparison techniques that the tool provides. Based on this, the tool retrieves a ranked list of ontologies for each criterion. Finally, a unique ranking is defined by means of a global aggregated measure which combines the different selected criteria, using rank fusion techniques [1].

Once the system retrieves those ontologies closely related to the Golden Standard, it supports an additional step in the evaluation process, by implementing a Collaborative Filtering approach [12][14][19]. Since some ontology features can only be assessed by humans, the last evaluation step takes into consideration the manual feedback provided by users of the ontologies, to re-rank the list of ontologies, thus retrieving not only the ontologies that best fit the Golden Standard, but the most qualified ones according to human evaluations.

The paper is organized by the following structure. In Section 2 we present the relevant work related to our research; section 3 describes the system architecture; sections 4 and 5 present the automatic evaluation measures used by the system; and some conclusions are given in section 6.

2. RELATED WORK
Our research addresses problems in three different research areas, where we draw from prior related work. These are: ontology evaluation and reuse, which is the primary goal of our work; rank fusion, which is used to combine the ratings provided by different ontology evaluation criteria; and collaborative filtering, by which we get further evaluation measures for ontology features that are better assessed by human judgment.

2.1 Ontology Evaluation
Different methodologies for ontology evaluation have been proposed in the literature considering the characteristics of the ontologies and the specific goals or tasks that the ontologies are intended for. An overview of ontology evaluation approaches is presented in [2], where four different categories are identified:

- Those that evaluate an ontology by comparing it to a Golden Standard, which may itself be an ontology [11] or some other kind of representation of the problem domain for which an appropriate ontology is needed.

- Those that evaluate the ontologies by plugging them in an application, and measuring the quality of the results that the application returns [16].
- Those that evaluate ontologies by comparing them to unstructured or informal data (e.g. text documents [3]) which represent the problem domain.
- Those based on human interaction to measure ontology features not recognizable by machines [10].

In each of the above approaches, a number of different evaluation levels might be considered to provide as much information as possible. Several levels can be identified in the literature:

- The lexical level [3][11][21] which measures the quality by comparing the words (lexical entries) of the ontology with a set of words that represent the problem domain.
- The taxonomy level [11] which considers the hierarchical connection between concepts using the *is-a* relation.
- Other semantic relations besides hierarchical ones [6][8].
- The syntactic level [7] which considers the syntactic requirements of the formal language used to describe the ontology.
- Context or application level [4] which considers the context of the ontology, such as the ontologies that reference or are referenced by the one being evaluated, or the application it is intended for.
- The structure, architecture and design levels [10] which take into account the principles and criteria involved in the ontology construction itself.

Table 1 summarizes all these approaches [2].

Table 1. An overview of approaches to ontology evaluation

Approach to evaluation				
Level	Golden Standard	Application based	Data Driven	Assessment by humans
Lexical entries, vocabulary, concept, data	X	X	X	X
Hierarchy, taxonomy	X	X	X	X
Other semantic relations	X	X	X	X
Context, application		X		X
Syntactic	X			X
Structure, architecture, design				X

In the present paper, two novel evaluation measures are proposed. The first one is based on a Golden Standard approach and the lexical level measure proposed by Maedche and Staab [11]. The second one is based on assessment by humans in a collaborative filtering approach.

2.2 Rank Fusion

Rank fusion has been a widely addressed research topic in the field of Information Retrieval [1][5][9]. Given a set of rankings which apply to a common universe of information objects, the task of rank aggregation consists of combining this list in a way to optimize the performance of the combination. Examples where rank fusion takes place include, for instance, metasearch [18] distributed search from heterogeneous sources, personalized retrieval, classification based on multiple evidence, etc.

Fusion techniques typically bring better recall, better precision, and more consistent performance than the individual systems being combined [1]. Fusion techniques can be characterized by:

- The input data they require: ranks, scores, or full information of the objects.
- Whether or not training data is used, which usually consists of manual relevance judgments on the information objects.
- The degree of overlap between the sets of rated objects, ranging from total overlap (a.k.a. data fusion), to completely disjoint sets (a.k.a. collection fusion), and arbitrarily overlapping.
- The application level, which can be a) external, if autonomous rating systems are integrated into a new meta layer, or b) internal, if the combination takes place at heart of a retrieval system, where different subsystems collect evidence from several sources or different criteria.

In our work, rank fusion techniques are used to combine the individual ontology lists retrieved by partial evaluation criterion into an aggregated ontology ranking. This can be understood as a metasearch problem where a) the input data are the evaluation ratings from different criteria, b) no training data is used (there are no prior manual rating or reference judgements for comparison), c) the overlap is complete (all the evaluation criteria are applied on the same ontology repository), and d) the level of application is internal (the rating sources are components within the CORE system).

2.3 Collaborative Filtering

Collaborative filtering strategies make automatic predictions (filter) about the interests of a user by collecting taste information from many users (collaborating). This approach usually consists of two steps: 1) look for users that have a similar rating pattern to that of the active user (the user for whom the prediction is done), and 2) use the ratings of users found in step 1 to compute the predictions for the active user. These predictions are specific to the user, differently to those given by more simple approaches that provide average scores for each item of interest, for example based on its number of votes.

Collaborative filtering is a widely explored field. Three main aspects typically distinguish the different techniques reported in the literature [14]: user profile representation and management, filtering method, and matching method.

User profile representation and management can be divided into five different tasks:

- Profile representation. Accurate profiles are vital for the content-based component (to ensure recommendations are appropriate) and the collaborative component (to ensure

that users with similar profiles are in fact similar). The type of profile chosen in this work is the user-item ratings matrix (ontology evaluations based on specific criteria).

- Initial profile generation. The user is not usually willing to spend too much time in defining her/his interests to create a personal profile. Moreover, user interests may change dynamically over time. The type of initial profile generation chosen in this work is a manual selection of values for only five specific evaluation criteria.
- Profile learning. User profiles can be learned or updated using different sources of information that are potentially representative of user interests. In our work, profile learning techniques are not used.
- The source of user input and feedback to infer user interests from. Information used to update user profiles can be obtained in two different ways: using information explicitly provided by the user, and using information implicit observed in the user's interaction. Our system uses no feedback to update the user profiles.
- Profile adaptation. Techniques are needed to adapt the user profile to new interests and forget old ones as user interests evolve with time. Again, in our approach profile adaptation is done manually (manual update of ontology evaluations).

Filtering method. Products or actions are recommended to a user taking into account the available information (items and profiles). There are three main information filtering approaches for making recommendations:

- Demographic filtering: Descriptions of people (e.g. age, gender, etc) are used to learn the relationship between a single item and the type of people who like it.
- Content-based filtering: The user is recommended items based on the descriptions of items previously evaluated by other users. Content-based filtering is chosen approach in our work (the system recommends ontologies using previous evaluations of those ontologies).
- Collaborative filtering: People with similar interests are matched and then recommendations are made.

Matching method. Defines how user interests and items are compared. Two main approaches can be identified:

- User profile matching: People with similar interests are matched before making recommendations.
- User profile-item matching: A direct comparison is made between the user profile and the items. The degree of appropriateness of the ontologies is computed by taking into account previous evaluations of those ontologies.

In CORE, a new ontology evaluation measure based on collaborative filtering is proposed, considering user's interest and previous assessments of the ontologies.

3. SYSTEM ARCHITECTURE

In this section we describe the architecture of CORE, our Collaborative Ontology Reuse and Evaluation environment. Figure 1 shows the overview of the system. We distinguish three different modules. The first one, the left module, receives the Golden Standard definition as a set of initial terms and allows the user to modify and extend it using WordNet [13]. The second one, represented in the center of the figure, allows the user to select a set of ontology evaluation techniques provided by the system to recover the ontologies closest to the given Golden Standard. The third one, or right one, is a collaborative module that re-ranks the list of recovered ontologies, taking into consideration previous feedback and evaluations of the users.

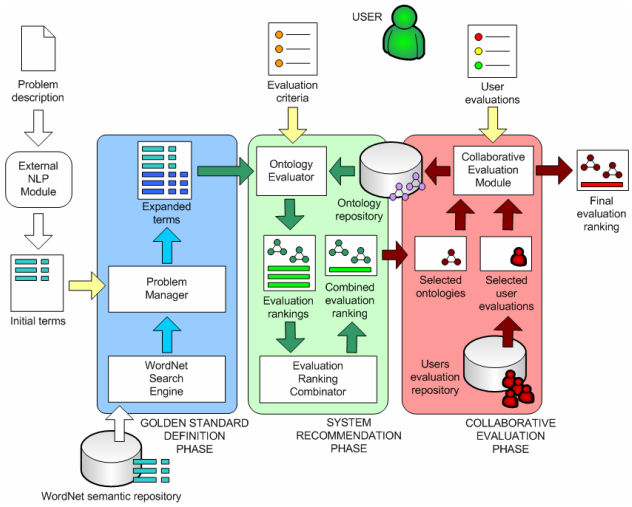

Figure 1. CORE architecture

3.1 Golden Standard Definition Phase

The Golden Standard Definition module receives an initial set of terms. These terms are supposed to be obtained by an external Natural Language Processing (NLP) module from a set of documents related to the specific domain in which the user is interested. This NLP module would receive the repository of documents and return a list of pairs (lexical entry, part of speech), that roughly represents the domain of the problem. This phase is part of future work. Here in our experiments, the list of initial (root) terms has been manually assigned.

The module allows the user to expand the root terms using WordNet [13] and some of the relations it provides: hypernym, hyponym and synonym. The new terms added to the Golden Standard using these relations might also be extended again and added to the problem definition.

The final representation of the Golden Standard can be defined as a set of terms $T(L^G, POS, L^{GP}, R, Z)$ where:

- L^G is the set of lexical entries defined for the Golden Standard.
- POS corresponds to the different Parts Of Speech considered by WordNet: *noun*, *adjective*, *verb* and *adverb*.
- L^{GP} is the set of lexical entries of the Golden Standard that have been extended.
- R is the set of relations between terms of the Golden Standard: *synonym*, *hypernym*, *hyponym* and *root* (if a term has not been obtained by expansion, but is one of the initial terms).
- Z is an integer number that represents the depth or distance of a term to the root term from which it has been derived.

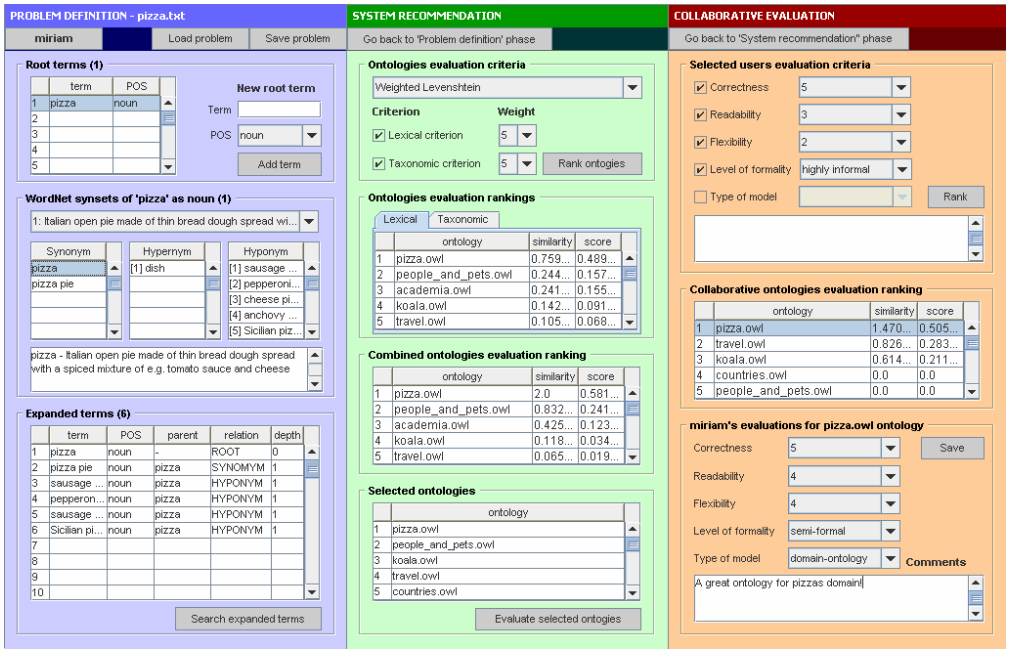

Figure 2. CORE user interface

Example:

T_1 ("pizza", noun, "", ROOT, 0). T_1 is one of the root terms of the Golden Standard. The lexical entry that it represents is "pizza", its part of speech is "noun", it has not been expanded from any other term so its lexical parent is the empty string, its relation is ROOT and its depth is 0.

T_2 ("pizza pie", noun, "pizza", Synonym, 1). T_2 is a term expanded from T_1. The lexical entry it represents is "pizza pie", its part of speech is "noun", the lexical entry of its parent is "pizza", it has been expanded by the synonym relation and the number of relations that separated it from the root term T_1 is 1.

The left part of Figure 2 shows the interface of the Golden Standard Definition phase. In the top level we can see the list of root terms. The user is allowed to manually insert new root terms giving their lexical entries and selecting their parts of speech. The correctness of these new insertions is controlled by verifying that all the considered lexical entries belong to the WordNet [13] repository. In the bottom level we can see the final Golden Standard definition: the final list of (root and expanded) terms that represent the domain of the problem. In the intermediate level it can be seen how the user can make a term expansion. The user selects one of the previous terms from the Golden Standard definition and the system shows him all its meanings contained in WordNet [13]. After he chooses one, the system automatically presents him three different lists with the synonyms, hyponyms and hypernyms of the term. The user can choose one or more elements of these lists and they will automatically be added to the expanded terms list. For each expansion the depth of the new term is increased by one unit. This will be used later to measure the importance of the term within the Golden Standard: the greater the depth of the derived term with respect to its root term, the less its relevance will be.

3.2 System Recommendation Phase

In this phase the system should retrieve the ontologies that better conceptualize the Golden Standard domain. The middle module of Figure 1 represents the structure of the recommendation phase of the system. Firstly, the user selects a set of evaluation criteria to be performed. After considering the selected criteria and taking into account the Golden Standard and the ontologies of the repository, the system retrieves a ranked list of ontologies (ordered by their similarity to the Golden Standard) for each criterion. Then, all these lists are merged using rank fusion techniques [1] to obtain a global measure.

The middle part of Figure 2 represents the user interface of the System Recommendation module. In the upper level we distinguish the criteria selection phase. By now, two content evaluation criteria can be selected to retrieve the most similar ontologies: 1) the lexical criterion, which measures similarity between the lexical entries of the Golden Standard and the lexical entries of the ontologies and, 2) the taxonomic criterion, which evaluates the hierarchical structure between them. The user can also select the relevance of each criterion in the rank aggregation process, using a range of discrete values [1, 2, 3, 4, 5], where 1 symbolizes the lowest relevance value and 5 the highest. Moreover, different kinds of lexical and taxonomic similarity measures have been implemented and tested using this tool. These may now be selected in this phase. These measures will be explained in section 4 of this document. The intermediate level of Figure 2 shows a different ranked list for each criterion and the final fused list. In each of these tables, two different ratings are displayed for each ontology. The first one refers to the similarity between the ontology and the Golden Standard. The second rating, score, shows the similarity value normalized by the sum of all the values. The score measure exhibits the distribution of the ratings and allows us to better evaluate the different techniques.

Once the final ranked list has been retrieved, the system allows the user to select a subset of ontologies that he considers adequate for the Collaborative Evaluation Phase.

3.3 Collaborative Evaluation Phase

This module has been designed to confront the challenge of evaluating those ontology features that are by their nature, more difficult for machines to address. Where human judgment is required, the system will attempt to take advantage of Collaborative Filtering techniques [12][14][19]. Some approaches for ontology development [20] have been presented in the literature concerning collaboration techniques. However to our knowledge, Collaborative Filtering strategies have not yet been used in the context of ontology reuse.

The collaborative module ranks and presents the best ontologies for the user, taking into consideration previous manual evaluations.

Several issues have to be considered in a collaborative system. The first one is the representation of user profiles. The type of user profile selected for our system is a user-item rating matrix *(ontologies evaluations based on specific criteria)*. The initial profile is designed as a manual selection of five predefined criteria [15]:

- Correctness: specifies whether the information stored in the ontology is true, independently of the domain of interest.
- Readability: indicates the non-ambiguous interpretation of the meaning of the concept names.
- Flexibility: points out the adaptability or capability of the ontology to change.
- Level of Formality: highly informal, semi-informal, semi-formal, rigorously-formal.
- Type of model: upper-level (for ontologies describing general, domain-independent concepts), core-ontologies (for ontologies describing the most important concepts on a specific domain), domain-ontologies (for ontologies describing some domain of the world), task-ontologies (for ontologies describing generic types of tasks or activities) and application-ontologies (for ontologies describing some domain in an application-dependent manner).

The above criteria can be divided in two different groups: 1) the *discrete* criteria (correctness, readability and flexibility) that are represented by discrete numeric values [0, 1, 2, 3, 4, 5] where 0 indicates that the ontology does not fulfill the criterion, and 5 indicates the ontology completely satisfies the criterion and, 2) the *boolean* criteria (level of formality and type of model) are represented by a specific value that is either satisfied by the ontologies, or not.

The collaborative system does not implement any profile learning technique or relevance feedback to update user profiles. But, the profiles may be modified manually.

After the user profile has been defined, it is important to select an appropriate type of filtering. For this work, a content-based filtering technique has been chosen; this means, ontologies (our content items) are recommended based on previous users evaluations.

Finally, a type of matching must also be picked out for the recommendation process. In this work, a novel technique of User Profile-Item matching is proposed. To evaluate the levels of relevance of the ontologies, the technique will make comparisons between the user's interests and the ontology's evaluations stored into the system. This will be explained in section 5.

The right portion of Figure 2 shows the Collaborative Evaluation module. At the top level the user's interest can be selected as a subset of criteria with associated values representing thresholds that manual evaluations of the ontologies should fulfil. For example, when a user sets a value of 3 for the correctness criterion, the system recognizes that he is looking for ontologies whose correctness value is greater than or equal to 3. Once the user's interests have been defined, the set of manual evaluations stored in the system is used to compute which ontologies fit his interest best. The intermediate level shows the final ranked list of ontologies returned by the Collaborative Filtering module. To add new evaluations to the system, the user must select an ontology from the list and choose one of the predetermined values for each of the five aforementioned criteria. The system also allows the user to add some comments to the ontology evaluation in order to provide more feedback.

One more action has to be performed in order to visualize the evaluation results of a specific ontology. Figure 3 shows the user's evaluation module. On the left side we can see the summary that the system provides of the existing ontology evaluations with respect to the user's interest. In Figure 3, 3 of 6 evaluations of the ontology have fulfilled the correctness criteria, 5 of 6 evaluations have fulfilled the readability criteria, and so on. On the right side, we can see how the system enables the user to observe all the evaluations stored into the system about a specific ontology. This may be of interest since we may trust some users more than others during the Collaborative Filtering process.

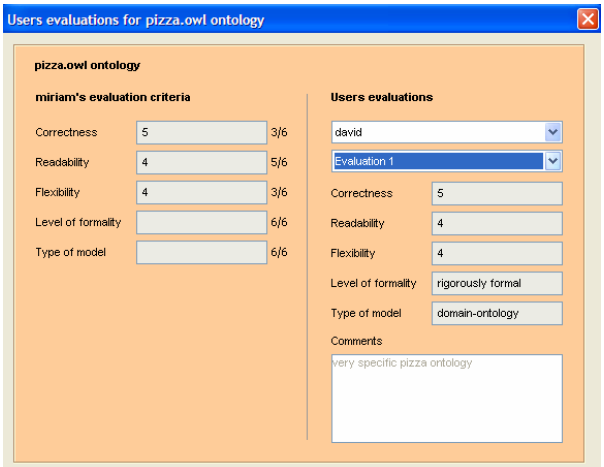

Figure 3. CORE user's evaluations

4. CONTENT ONTOLOGY EVALUATION

In order to obtain similarities between the Golden Standard and the stored ontologies, two different content ontology evaluation levels have been considered, the lexical and the taxonomic. Several measures have been developed and tested for each level.

In the following sections we present the approaches that have shown better performance.

4.1 Lexical Evaluation Measures

The lexical evaluation assesses the similarity between the domain of the problem as described by the Golden Standard and an ontology by comparing the lexical entries, or words that represent them. A new lexical evaluation measure based on Maetche and Staab work [11] will be explained in this section. Some definitions must first be introduced.

Definition 1 (Lexical entry). A lexical entry l represents a string or word.

Definition 2 (Golden Standard Lexicon). The Golden Standard Lexicon, L^G is defined by the set of lexical entries extracted from the terms of the Golden Standard, where each term has a single lexical entry that represents it.

Definition 3 (Ontology Lexicon). The Ontology Lexicon, L^O is defined as the set of lexical entries extracted from the Concepts of the Ontology. Each concept is represented by one or more lexical entries that are extracted from the concept name, the rdfs:label property value, or other properties that could be added to the lexical extraction process considering the characterization of each ontology.

Definition 4 (Levenshtein distance). The Levenshtein distance, $ed(l_i, l_j)$ between two lexical entries l_i and l_j measures the minimum number of token insertions, deletions and substitutions to transform l_i into l_j using a dynamic algorithm.

Example: $ed(\text{"pizzapie"}, \text{"pizza\_pie"}) = 1$

Maedche and Staab [11] propose a lexical similarity measure for strings called String Matching. This method compares two lexical entries l_i, taking into account the Levenshtein distance against the shortest lexical entry.

$$\text{SM}(l_i, l_j) = \max(0, \frac{\min(|l_i|, |l_j|) - ed(l_i, l_j)}{\min(|l_i|, |l_j|)}) \in [0,1]$$

SM returns a degree of similarity between 0 and 1, where 0 is a null match and 1 represents a perfect match.

Example: SM("pizzapie", "pizza_pie") = 7/8.

Based on the String Matching they propose a lexical similarity measure to compare an ontology to a Golden Standard, by computing the average string matching between the set of Golden Standard lexical entries and the set of ontology lexical entries:

$$\overline{\text{SM}}(L^G, L^O) = \frac{1}{|L^G|} \sum_{l_i \in L^G} \max_{l_j \in L^O} SM(l_i, l_j)$$

$\overline{\text{SM}}(L^G, L^O)$ is an asymmetric measure that determines the extent to which the lexical level of the Golden Standard is covered by the lexical level of the Ontology. Future work must be done in order to penalize those ontologies which contain all the strings of the Golden Standard but also many others.

There is one principle difference between that approach and ours; Maedche defines the Golden Standard as an ontology, while we use our own model. This fact provides us with the capability to use all the additional information stored in the Golden Standard in order to improve content evaluation measures.

When a domain is modeled as a set of lexical entries, some lexical entries have greater relevance when defining the semantics than do others. Assuming this characteristic we have decided to distinguish the importance of the Golden Standard terms. The root terms are considered the most representative ones while the relevance of the expanded terms depends on the number of relations that separate them from a root term. With this modification we emphasize the main semantics and relegate the complementary ones into the background. In this work we define the Golden Standard Lexical weight measure to evaluate the importance of each term.

Definition 5 (Golden Standard lexical weight). Given a list of lexical entries $L = \{l_i\}$ expanded from a common root lexical entry, we define the weight of $l \in L$ as:

$$w(l) = \begin{cases} 1 + \dfrac{\max_i(Depth(l_i)) - Depth(l)}{\max_i(Depth(l_i))} \in [1,2] & \text{if } |L| > 1 \\ 2 & otherwise \end{cases}$$

The value returned is represented as a degree of relevance between 1 (the farthest distance to the root lexical entry), and 2 (no distance to the root lexical entry). If the root lexical entry has not been expanded we assign it the weight value 2.

Figure 4 shows an example of this measure, where T_1 is the root term and consequently has the greater weight. T_3 is the most remote term and it has the smaller weight. The intermediate terms like T_2 have a weight between the maximum and the minimum relative to their distance from the root term.

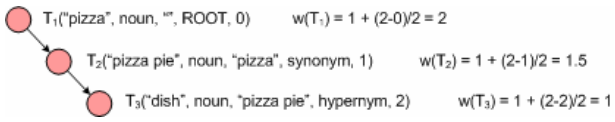

Figure 4. Golden Standard lexical weight measure

In our approach, we have modified the previous lexical measure taking into account the weight or relevance of each term to represent the semantics of the domain.

$$\overline{\text{SM}}(L_G, L_O) = \frac{1}{|L_G|} \sum_{li \in L_G} \max_{lj \in Lo} SM(l_i, l_j) \cdot w(l_i)$$

Through our experiments, this new measure has been shown to better discriminate the ontologies, giving a higher similarity value to the ontologies that are closer to the Golden Standard and lower rating to the ontologies that worst fit the problem domain. Future work is needed in order to give more or less relevance to the derived terms of the Golden Standard using not only their distances to the root terms but also, the kind of relation from which they have been derived, synonym, hypernym or hyponym.

4.2 Taxonomic Evaluation Measures

The taxonomic evaluation assesses the degree of overlapping between the hierarchical structure of the ontology, defined by the "is-a" relation and the Golden Standard structure, defined by the derivations of terms to complete the domain representation. The following notations and definitions will be used to define our measure:

$T_i^G \in T^G$ represents a Golden Standard term.

$C_i^O \in C^O$ represents an Ontology concept.

Definition 8 (Semantic Cotopy of a Golden Standard Term). The Semantic Cotopy of a Golden Standard term $SC(T_i^G)$ is defined as the set of lexical entries of the terms derived from the same root term as T_i^G, including the lexical entries of T_i^G.

Definition 9 (Semantic Cotopy of an Ontology Concept).

The Semantic Cotopy of an Ontology concept $SC(C_i^O)$ is defined as the set of lexical entries of the concepts related with C_i^O in the ontology with a direct relation, including the lexical entries of C_i^O.

Given Maedche and Staab [11] measure, an adaptation is performed to obtain a similarity between an ontology Concept and a Golden Standard Term relative to the new Golden Standard definition. The similarity is computed as the intersection between the Semantic Cotopy of the Golden Standard term and the Semantic Cotopy of the ontology concept normalized by the total possible overlap.

$$TS(T_i^G, C_i^O) = \frac{|SC(T_i^G) \cap SC(C_i^O)|}{|SC(T_i^G) \cup SC(C_i^O)|}$$

In order for two lexical entries to be considered a match, their similarity must be greater than a threshold empirically estimated as 0.2. For similarities below this value we have observed there is no significant morphological resemblance between terms.

The taxonomic similarity measure considers all the overlaps between the Ontology and the Golden Standard. In order to optimize the evaluation, only a subset of terms and concepts are used to assess the taxonomic similarity. This subset is obtained through the lexical measurement, this is done by selecting only the terms and concepts that have matched with a similarity value greater than 0.2.

$$\overline{TS}(T^G, C^O) = \frac{1}{|T^G|} \sum_{\substack{T_i^G \in T^G, C_j^O \in C^O \wedge \\ \exists l_i \in T_i^G, \exists l_j \in C_j^O : SM(l_i, l_j) > 0.2}} TS(T_i^G, C_j^O)$$

5. COLLABORATIVE FILTERING FOR ONTOLOGY REUSE

In this section, a new automatic evaluation measure that exploits the advantages of Collaborative filtering is proposed. It will match the set of ontologies or items that better fulfill the user's interest exploring the set of manual evaluation stored into the system. As we explained in section 3 user's evaluations are represented like a set of five defined criteria and their respective values manually determined by the user who makes the evaluation. On the other hand, user's interests are expressed like a subset of those criteria, and their respective values, meaning a threshold or restriction to be satisfied by user's evaluations.

Two main steps are presented for this measure. The first one describes how the similarity degree between a user's evaluation criterion and a user's interest threshold for the same criterion is assessed. The second one describes how calculate the final rankings of the ontologies.

5.1 Collaborative Evaluation Measures

As we explained in section 3.3, the user evaluation about a specific ontology is made considering five different criteria. These five criteria are divided in two different groups: 1) the *discrete* criteria (correctness, readability and flexibility), which take discrete numeric value [0, 1, 2, 3, 4, 5], where 0 means that the ontology does not fulfill the criterion, and 5 means the ontology completely satisfy the criterion, and 2) the *boolean* criteria (level of formality and type of model) that are represented by specific values that can be or not satisfied by the ontology. User's interests are defined like a subset of those criteria and their respective values representing a set of thresholds that the ontologies should fulfill. The user's interests are sized up against the respective values of those criteria in the user's evaluations or user's profiles.

For the boolean case, a value of 0 is returned if the value of the criterion *n* in the evaluation *m* does not fulfill the user's requirements for that criterion, and 2 otherwise.

$$similarity_{bool}(criterion_{mn}) = \begin{cases} 0 & \text{if } evaluation_{mn} \neq threshold_{mn} \\ 2 & \text{if } evaluation_{mn} = threshold_{mn} \end{cases}$$

For the discrete case, the measure includes different aspects: a similarity assessment and a penalty assessment. The similarity assessment is based on the distance between the value of the criterion n into the evaluation m, and the threshold specified in the user's interest for that criterion. The more the value of the criterion n in evaluation m overcomes the threshold specified for this criterion, the greater the similarity value is. The penalty assess considers how difficult is to surpass this threshold. The more difficult to surpass the threshold, the lower the penalty value is.

$$similarity_{num}(criterion_{mn}) = \\ = 1 + similarity^*_{num}(criterion_{mn}) \cdot penalty_{num}(threshold)$$

This measure also returns values between 0 and 2. The consideration of retrieving a similarity value between 0 and 2 has taken from other collaborative matching measures [19] to not manage negative numbers. In the case of this collaborative measure, negative similarity values would be returned when the value of the criterion in the user's evaluations does not surpass the threshold required in the user's interests.

5.2 Collaborative Evaluation Ranking

The user's interests and the user's profiles, or evaluations of the ontologies stored into the system, are used to make the final ranking of the ontologies. The similarity between an ontology evaluation and the user's requirements is measured as the average of its *N* criteria similarities.

$$similarity(evaluation_m) = \frac{1}{N} \sum_{n=1}^{N} similarity(criterion_{mn})$$

The similarity of a specific ontology to the user's requirements is measured as the average of the *M* evaluations similarities for that specific ontology.

$$similarity(ontology) = \frac{1}{M}\sum_{m=1}^{M} similarity(evaluation_m)$$
$$= \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N} similarity(criterion_{mn})$$

In case of ties, the final collaborative ranking sorts the ontologies taking into account not only the average similarity between the ontologies and the evaluations stored into the system, but also the total number of evaluations of those ontologies, providing more relevance to those ontologies that have been rated more times.

$$\frac{M}{M_{total}} similarity(ontology)$$

6. CONCLUSIONS AND FUTURE WORK

In this work a new tool for ontology evaluation and reuse have been presented, including some interesting features like a new Golden Standard model, new lexical evaluation criteria, the use of rank fusion techniques to combine different content ontology evaluation measures, and the use of a novel Collaborative filtering strategy to take advantage of user's opinions in order to automatically evaluate features that only can be assessed by humans.

Some initial experiments, not explained in this paper, have been developed using a set of ontologies form the Protégé OWL repository [17] obtaining positive results, but a more detailed and rigorously experimentation must be done in order to achieve relevant conclusions.

7. ACKNOWLEDGMENTS

Thanks to Enrico Motta, the "great director" of the SSSW2005, Aldo Gangemi and Elke Michlmayr, also members of this fantastic School, for our 2 a.m. conversations about ontology evaluation. Thanks to Alexander Gomperts for his help reviewing the paper, and special gratefulness to Denny Vrandecic for encourage us to do this work. This research was supported by the Spanish Ministry of Science and Education (TIN2005-0685).

8. REFERENCES

[1] Aslam, J. A., Montague, M. Models for metasearch. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001). New Orleans, Louisiana, 2001, pp. 276-284.

[2] Brank J., Grobelnik M., Mladenić D. A survey of ontology evaluation techniques. SIKDD 2005 at multiconference IS 2005, 17 Oct 2005, Ljubljana, Slovenia.

[3] Brewester, C. et al. Data driven ontology evaluation. Proceedings of Int. Conf. on Language Resources and Evaluation, Lisbon, 2004.

[4] Ding, L., et al., Swoogle: A search and metadata engine for the semantic web. Proc. CIKM 2004, pp. 652–659.

[5] Fox, E. A., Koushik, M. P., Shaw, J., Modlin, R., Rao, D. Combining evidence from multiple searches. 1st Text REtrieval Conference (TREC 1). Gaithersburg, Maryland, March 1992, pp. 319-328.

[6] Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J. A Theoretical Framework for Ontology Evaluation and Validation. In Proceedings of SWAP2005.

[7] Gomez-Perez, A. Some Ideas and Examples to Evaluate Ontologies. In Proceeding of the 11th Conference on Artificial Intelligence. Los Angeles, CA, February, pp. 299-305, 1995.

[8] Guarino, N., Welty, C. Evaluating Ontological Decisions with OntoClean. Communications of the ACM. 45(2):61-65. New York: ACM Press, 2002.

[9] Lee, J. H. Analyses of multiple evidence combination. 20th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 97). New York, 1997, pp. 267-276.

[10] Lozano-Tello, A., Gómez-Pérez, A. Ontometric: A method to choose the appropriate ontology. J. Datab. Mgmt., 15(2):1–18 (2004).

[11] Maedche, A., Staab, S. Measuring similarity between ontologies. Proc. CIKM 2002. LNAI vol. 2473, pp. 251-263.

[12] Masthoff J. Group modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. User Modeling and User-Adapted Interaction 14: 37-85, 2004.

[13] Miller, G. WordNet: A lexical database. Communications of the ACM, 38(11):39-41. 1995.

[14] Montaner M., López B., De la Rosa J.L. A Taxonomy of Recommended Agents on the Internet. Artificial intelligence Review 19: 285-330, 2003.

[15] Paslaru, E. Using Context Information to Improve Ontology Reuse. Doctoral Workshop at the 17th Conference on Advanced Information Systems Engineering CAiSE'05.

[16] Porzel, R., Malaka, R. A task-based approach for ontology evaluation. ECAI 2004 Workshop Ont. Learning and Population.

[17] Protégé OWL ontology Repository.

protege.stanford.edu/plugins/owl/owl-library/index.html

[18] Renda, M. E., Straccia, U. Web metasearch: rank vs. score based rank aggregation methods. ACM symposium on Applied Computing. Melbourne, Florida, 2003, pp. 841-846.

[19] Resnick P. et al. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. Internal Research Report, MIT Center for Coordination Science, March 1994.

[20] Sure Y., Erdmann M., Angele J., Staab S., Studer R., Wenke D. "OntoEdit: Collaborative Ontology Development for the Semantic Web". ISWC 2002.

[21] Velardi, P., et al. Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. In Ont. Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.