# Dynamic Integration of Multiple Evidence Sources for Ontology Learning

Gerhard Wohlgenannt[1], Albert Weichselbraun[2], Arno Scharl[3], Marta Sabou[3]

[1] Vienna University of Economics and Business, Augasse 2-6, 1090 Wien, Austria
gerhard.wohlgenannt@wu.ac.at
[2] University of Applied Sciences Chur, Ringstrasse 34, 7004 Chur, Switzerland
albert.weichselbraun@htwchur.ch
[3] MODUL University Vienna, Am Kahlenberg 1, 1190 Wien, Austria
{arno.scharl, marta.sabou}@modul.ac.at

**Abstract.** Although ontologies are central to the Semantic Web, current ontology learning methods primarily make use of a single evidence source and are agnostic in their internal representations to the evolution of ontology knowledge. This article presents a continuous ontology learning framework that overcomes these shortcomings by integrating evidence from multiple, heterogeneous sources (unstructured, structured, social) in a consistent model, and by providing mechanisms for the fine-grained tracing of the evolution of domain ontologies. The presented framework supports a tight integration of human and machine computation. Crowdsourcing in the tradition of games with a purpose performs the evaluation of the learned ontologies and facilitates the automatic optimization of learning algorithms.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: Evidence Integration, Games with a Purpose, Knowledge Evolution, Ontology Learning

## 1. INTRODUCTION

Ontologies are formal conceptualizations of application domains [Gruber 1995] and provide a common understanding of domain concepts and relations among different stakeholder groups. Central to the architecture of the Semantic Web, ontologies facilitate text understanding and automatic processing of textual resources, and enable various applications such as question answering and information retrieval [Navigli et al. 2011]. The manual construction of ontologies, however, is a costly and cumbersome process - especially when ongoing updates and refinements are necessary to keep track of evolving domain knowledge. To improve productivity and reduce the human input required, a number of (semi-)automated ontology learning approaches emerged over the last decade focusing on tasks such as the learning of synonyms, concepts, taxonomies, relations and axioms.

Current ontology learning approaches tend to focus on extraction from a single type of data source, primarily text. A major drawback of using solely textual sources for ontology learning and evolution tasks is that many documents do not refer to important entities explicitly. These entities represent the common ground between readers and authors of a given community, so referring to them in an

The work presented in this article was developed within the DIVINE research project (www.weblyzard.com/divine), funded by the Austrian Ministry of Transport, Innovation and Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT Semantic Systems (www.ffg.at/fit-it). The authors would like to thank Michael Föls for his efforts in implementing and evaluating the social media application framework underlying the Climate Quiz, which was developed as part of Triple-C (www.ecoresearch.net/triple-c), a research project funded by the Austrian Climate Research Program of the Austrian Climate and Energy Fund (klimafonds.gv.at).

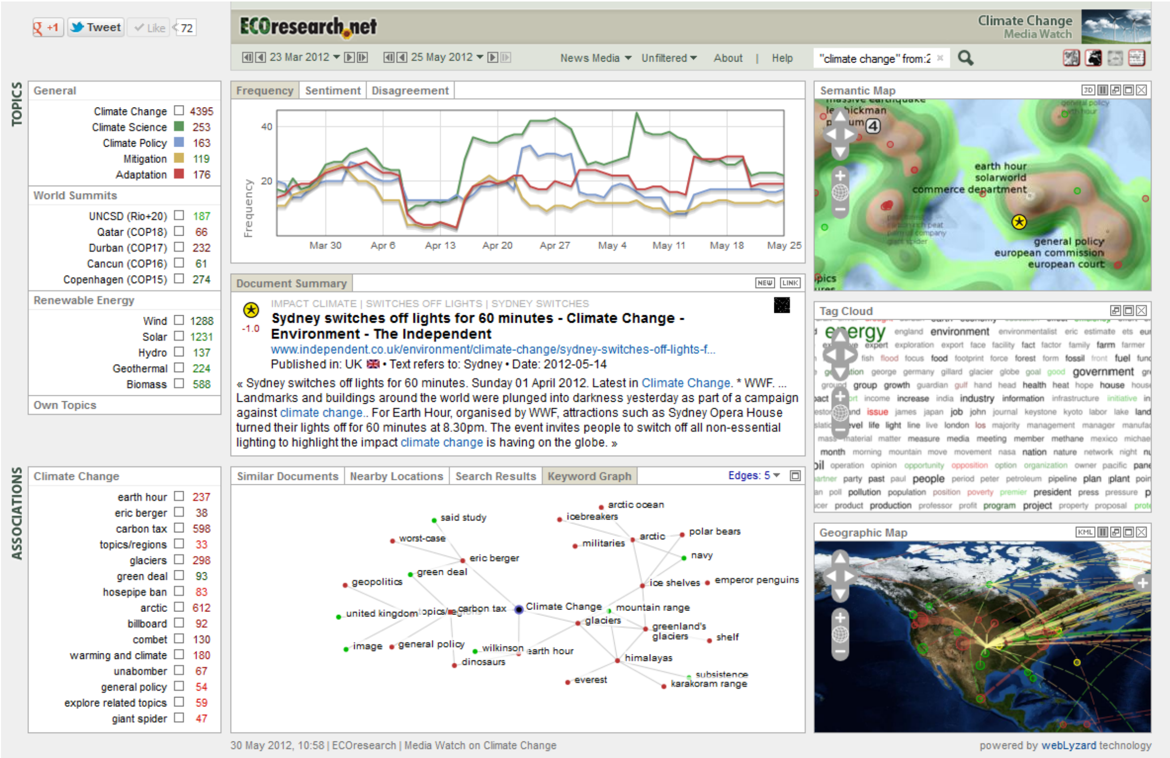


Fig. 1. Visual Dashboard of the Media Watch on Climate Change.

explicit manner is usually not necessary. Non-textual resources such as online ontologies, linked data (structured) and collective intelligence (social) in the form of folksonomies [Specia and Motta 2007] represent rich sources of complementary data. Furthermore, due to their high update frequency, social sources tend to reflect the latest terminology within a domain [Angeletou et al. 2007; Mika 2007]. The ontology learning framework described in this article, by contrast, makes use of and integrates multiple and evolving evidence sources. Integrating unstructured, structured and social sources bridges the gap between knowledge expressed in textual form, and knowledge captured in formal ontologies.

Another characteristic of ontology learning approaches is that they are generally designed for knowledge snapshots, where an ontology is learned at a certain point in time, without keeping track of its earlier versions. Through this, important insights into the evolution of the ontology are lost, especially when they are based on dynamically changing evidence sources (e.g., news or social media). Our aim is to support a continuous ontology learning process, where ontologies are updated in short and regular intervals. To that end, we use specialized data structures such as confidence matrices and source impact vectors (see Section 3) to track all changes over time in a fine-grained manner.

In current practice, the human feedback on the learned ontologies is primarily performed during ontology evaluation phases that are disconnected from the ontology learning process itself. This feedback only finds its way back into the learning process through the developer of the system, who evaluates the feedback and changes the system accordingly. We propose an architecture which tightly integrates the ontology learning algorithm with the human feedback stage. Namely, we crowdsource the evaluation of the extracted knowledge through (a set of) games with a purpose and immediately feed back this evaluation into the learning algorithm, thus efficiently and dynamically combining human and machine computation.

Section 2 provides an overview of related research. Section 3 then outlines the proposed method in detail, before Sections 4 and 5 detail how various evidence sources are harvested and integrated. The presented examples stem from the content repository of the Media Watch on Climate Change, a news and social media aggregator on climate change and related environmental issues. Its visual dashboard shown in Figure 1 provides access to the publications and postings of news media sites, blogs, Web 2.0 platforms (Facebook, Twitter, Google+, YouTube), environmental organizations, and Fortune 1000 companies. Section 6 describes the process of combining human feedback with machine learning, using a crowdsourcing approach in the tradition of games with a purpose to validate relations extracted from the Media Watch on Climate Change. Section 7 summarizes and concludes the article.

## 2. RELATED WORK

Our work primarily contributes to the research area of ontology learning, while also being related to ontology dynamics and crowdsourcing.

Up until now, learning lightweight ontologies has focused on techniques from corpus linguistics to extract semantically similar terms to form clusters of meaning [Wohlgenannt et al. 2009]. In recent years, integrated approaches have emerged, e.g. similarity computations or heuristics such as the lexico-syntactic patterns inspired by Hearst [Hearst 1992]. However, only a few initiatives [Correndo et al. 2009], [Weichselbraun et al. 2011] have started to exploit the increased availability of structured and social data sources [Sanchez and Moreno 2008]. As a result, ontology learning lags behind other fields such as ontology matching and ontology-based question answering where encouraging results were obtained by reusing structured third-party evidence [d'Aquin et al. 2008]. Similarly to ontology learning, the majority of ontology evolution approaches such as EVOLVA [Zablith et al. 2010], SPRAT [Maynard et al. 2009] and FLOR [d'Aquin et al. 2008] focus on a single source of data to derive changes, most frequently on text-based sources. With the exception of initial efforts to integrate evidence from multiple sources within the RELExO framework [Maynard and Aswani 2010], handling evolution triggered by multiple sources concurrently is an open research question in this field.

We aim to solve multi-source data integration and subsequent trend detection by using a confidence matrix-based approach. Confidences were already suggested in Text2Onto, which stores the learned ontology as a Probabilistic Ontology Model [Cimiano et al. 2009], and annotates every axiom with the confidence of the learning algorithm that generated it. While this work already suggested associating confidences to the learned axioms, it did not address the implications and possibilities related to ontology evolution. In the linked data community, dataset dynamics is a novel research trend, where the aim of current efforts is to monitor how the entire linked data cloud is evolving [Popitsch and Haslhofer 2010; Umbrich et al. 2010]. In contrast, we focus only on a selection of linked data sources relevant for a given application. We are interested not so much in the general evolution of data, but rather in a domain-specific interpretation of this evolution.

Games with a Purpose (GWAP) [Ahn and Dabbish 2008] leverage the participants' desire to be entertained in order to solve computational problems that are easy for humans but cannot yet be solved by computers reliably. Early attempts of using the GWAP paradigm in the area of the Semantic Web have primarily focused on tasks such as ontology learning and matching [Siorpaes and Hepp 2008a]. For ontology learning, OntoPronto [Siorpaes and Hepp 2008b] aims to build domain ontologies from Wikipedia articles by mapping these articles to the most specific class in the Proton ontology. Other games that aim to acquire semantic networks of common knowledge are the Virtual Pet Game [Kuo et al. 2009] and the Raport Game [Kuo et al. 2009].

A current trend is building games that make use of the large body of linked data sources in order to build new knowledge artifacts (Guess What ?! [Markotschi and Voelker 2010]), and to improve the quality of LOD sources (BetterRelation [Hees et al. 2011], WhoKnows? [Waitelonis et al. 2011], RISQ [Wolf et al. 2011]). The Guess What ?! game [Markotschi and Voelker 2010] creates ontologies containing complex class expressions by exploring instance data available as linked open data. Given a

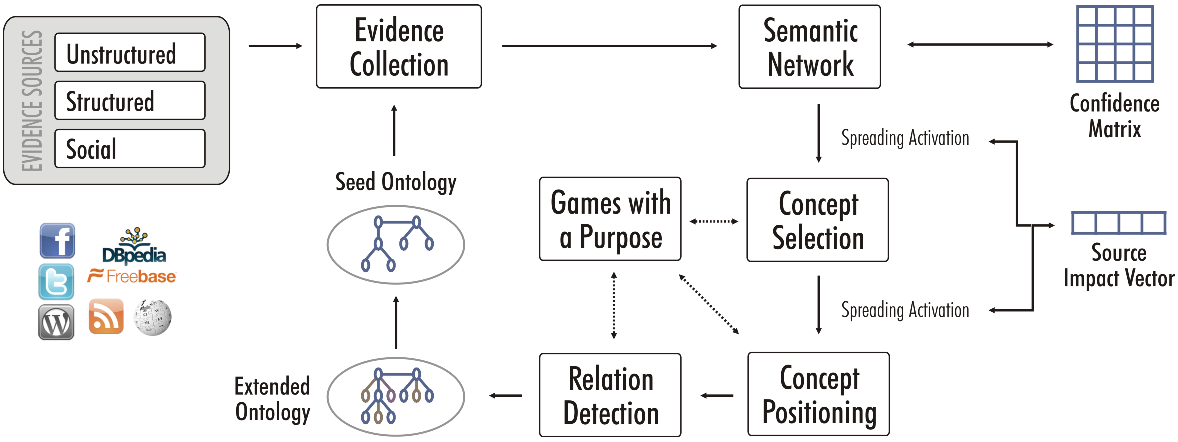


Fig. 2. Ontology Extension Architecture System Diagram.

seed concept (e.g., banana), the game collects relevant instances from DBpedia, Freebase and OpenCyc and extracts the main features of the concept (e.g., fruit, yellowish). In a first phase of a game, players must provide (i.e., guess), a suitable label for a set of attributes (e.g., banana and lemon). In a second phase, users verify each others responses and in the case of disagreement must specify the attribute they disagree upon (thus creating more detailed semantic descriptions, for example, that cherry is a fruit but not yellowish). Similarly, we incorporate and dynamically allocate multiple tasks (e.g., for verifying domain relevance of extracted relations) across different games. The novelty of this approach lies in the tight integration between the ontology learning algorithm and the games used, where multiple feedback cycles might take place. Current games do not exploit such a close link between human and machine computation.

## 3. METHOD OVERVIEW

This article presents an ontology learning framework built on top of work first presented in [Liu et al. 2005]. Since 2005, the architecture underwent various improvements, e.g. support for the detection of non-taxonomic relations [Weichselbraun et al. 2010b] or the integration of social media evidence sources [Weichselbraun et al. 2010a]. The ontology learning process (see Figure 2) can be summarized as follows: Starting from a typically small seed ontology, the system gains evidence for the inclusion of additional concepts from a heterogeneous set of resources with a variety of methods. All this evidence, which basically consists of *typed relations* between existing concepts and candidate concepts, is collected into a graph data structure which we call the *semantic network*. The confidence values from this semantic network are also stored in *confidence matrices*. In short, a confidence matrix represents the relation between two concepts from the ontology (one of the concepts is a candidate) and their connection strength observed in the evidence sources – as well as the dimension of time.

After the semantic network is filled with all evidence data, it is transformed into a spreading activation network. A source impact vector, which contains the impact values of the evidence source (a quality measure that reflects the trust in this particular source), contributes to setting the link weights in the spreading activation network. This allows the spreading activation process to detect new concepts and, in a further step, to identify the position of the concepts in the ontology, i.e. to which seed concept a new concept is connected to. In short, spreading activation facilitates the integration of the vast number of candidate concepts into a weighted network structure, the selection of a predefined number of new concepts, and the positioning of those new concepts in the ontology.

The evaluation of newly learned elements in the ontology is accomplished by using a set of Facebook-based games. Players of these games confirm (i) if a new concept is relevant for the domain, (ii) if the new concept has been connected to the correct seed concept, and (iii) eventually, suggest different relation types (e.g. isA, partOf, ...) for the relation between seed and candidate concept. Based on the feedback gathered from the game, and relying on the data from the confidence matrix, the system adapts the source impact vector (i.e., it assesses the reliability of individual sources). As a result, sources which yield a high ratio of relevant information have more impact on the ontology than sources that deliver low-quality evidence.

Tracking the evolution of (i) the source impact vector, (ii) the confidence matrix, and (iii) the connection strength between concepts obtained from spreading activation provides a comprehensive set of data for the detection of trends in terms of source quality, and the evolution of the domain.

### 3.1 Ontology Based Navigation Use Case

Ontologies serve as an important navigational aid within the Media Watch on Climate Change. The currently used representation as shown in Figure 3, however, is a static structure that does not adapt to the continuously updated knowledge repository. To obtain a dynamic structure reflecting the content streams from various evidence sources, the base architecture of the existing ontology learning system [Weichselbraun et al. 2010b] manages evidence integration with the semantic network, and concept selection and positioning with the help of spreading activation networks. The framework presented in this article extends and improves this system with additional and more-fine grained evidences, with a dynamic model to adapt and optimize the impact of evidence sources based on games with a purpose, and with the confidence matrix to assess the confidence of all ontology elements at a given time, thereby supporting and enabling a wide range of ontology evolution experiments.

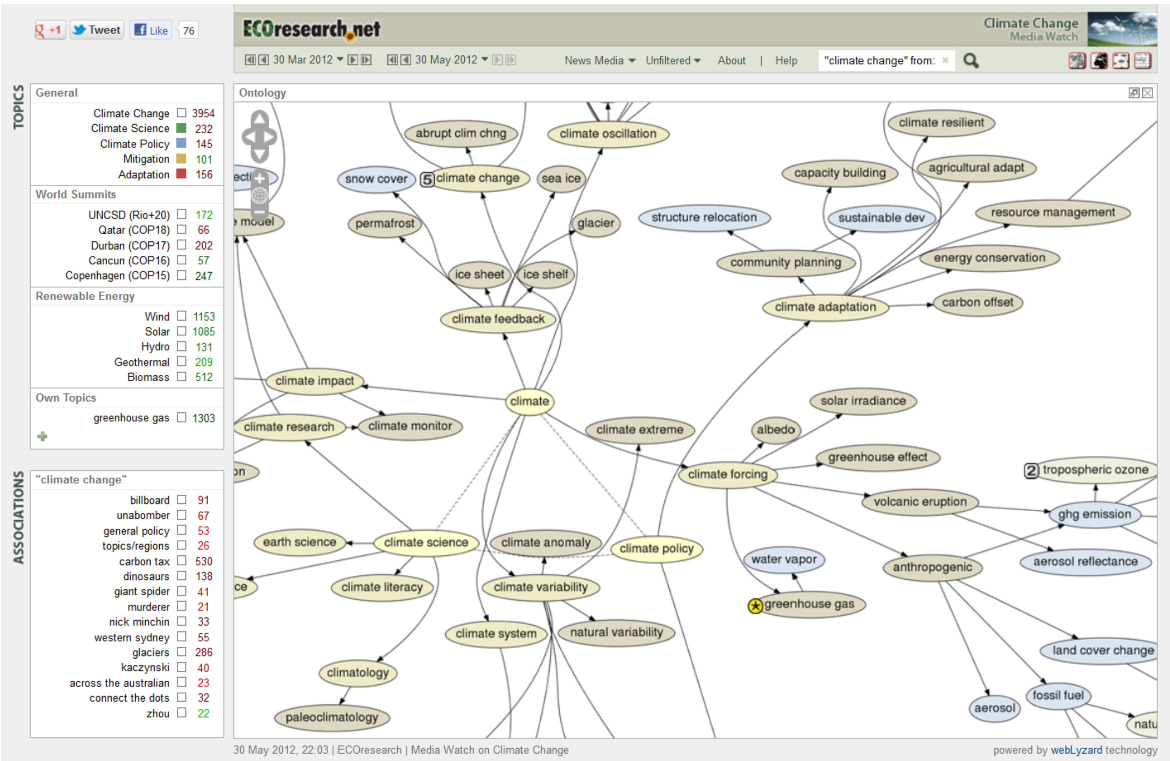


Fig. 3. Ontology Based Navigation in the Media Watch for Climate Change Portal.

## 4. COLLECTION OF EVIDENCE DATA

The first step of the ontology learning process is the collection of evidence data. The input to this process are *seed concepts.* For every seed concept $C_s$ (or rather its label) a suite of algorithms compute related terms. A network data structure (the "semantic network") then represents the relations between seed and candidate terms $C_c$. Seed concepts are either the concepts of the initial seed ontology, or concepts learned in previous ontology extension steps. Every algorithm that generates related terms for input concepts constitutes an evidence source. The following subsections classify the evidence sources used by the type of underlying dataset (text, social, structured).

### 4.1 Evidence from Text

Ontology learning from text relies on a (domain-specific) text corpus. As our goal is to enable monitoring of ontology evolution, the system ensures that the documents in those domain text corpora stem from the time interval under consideration (in most cases the last week or the last month). The authors have repeatedly and successfully applied the webLyzard suite of Web mining tools (www.weblyzard.com) for generating high-quality domain corpora with the required characteristics. Evidence sources based on temporally segmented domain corpora include [Liu et al. 2005]:

—*Keywords*: After compiling a target corpus (= set of documents or sentences where the term occurs) for each seed term and source, the system identifies keywords by comparing the term distribution of the target corpus with a reference corpus (all domain documents in the period) by means of co-occurrence statistics.

—*Trigger phrases*: Extract related terms scanning the domain text with Hearst-style patterns.

### 4.2 Evidence from Social Media

We distinguish two basic ways to access online social media for collecting evidence, i.e. typed relations to new term candidates, with seed terms as input:

—Direct access to *related terms* using the TagInfoService interface of the easy Web Retrieval Toolkit (www.semanticlab.net/index.php/eWRT). Currently the system collects evidence from Twitter, Flickr and Del.icio.us [Weichselbraun et al. 2010a].

—The webLyzard mirroring services contain components to query Youtube, Facebook, Twitter, etc. to generate domain text corpora. We then apply the extraction methods presented in the previous subsection to these corpora.

Weichselbraun et al. [Weichselbraun et al. 2010a] demonstrate the ontology learning system's benefit from the integration of terminology captured from online social media.

### 4.3 Evidence from Structured Data

The existing ontology learning system already uses WordNet as an evidence source and for disambiguation processes. The new system integrates additional structured data sources. The following structured sources serve as a starting point, with further sources being planned to be added over time:

—*DBpedia:* By leveraging components developed in [Weichselbraun et al. 2010b], the system queries the DBpedia SPARQL endpoint to gain new terms connected to a seed term. The query uses the seed term as subject, and properties such as *dcterms:subject* or *dbpedia-owl:wikiPageRedirects.* The resulting objects are candidates for the semantic network.

Table I. A simplified example of collected evidence for seed term *climate change*.

| seed ontology concept ($C_s$) | evidence source ($es$) | candidate concept ($C_c$) | description |
|---|---|---|---|
| climate change | oe:coOccurs | greenhouse gases | co-occurred in the NY Times |
| climate change | oe:twitterTag | petrol | related on Twitter |
| climate change | oe:dc-subject | copenhagen | connected with in DBpedia |
| climate change | ... | ... | ... |

—*Watson:* Watson [d'Aquin et al. 2007] is an ontology search engine, providing a single access point to a large collection of online available ontologies. Queries to the Watson API return concepts in online ontologies whose label matches the query term (the query term can be matched against concept labels or comments using either strict or partial matching techniques). Additional information about these concepts that is specified in online ontologies (e.g., alternative labels, more specific or more generic concepts) can then be used to extend the semantic network. Watson has been successfully used to support the reuse of knowledge from online ontologies [d'Aquin et al. 2008], [Sabou et al. 2007] for tasks as varried as open domain question answering [Lopez et al. 2010], folksonomy enrichment [Angeletou et al. 2007], ontology evolution [Zablith et al. 2010], relation detection [Sabou et al. 2008] and relation evaluation [Sabou et al. 2009].

—*Scarlet:* Scarlet [Sabou et al. 2008] is an extension over Watson that, given two input terms, identifies semantic relations between them that can be learned from the ontologies indexed by Watson. Unlike the previous sources that suggest related terms, Scarlet suggests typed relations between new and seed concepts. For example, when provided with two concepts labeled *Researcher* and *AcademicStaff*, Scarlet 1) identifies (at runtime) online ontologies that can provide information about how these two concepts inter-relate and then 2) combines this information to infer the relevant relation. The relation can be either provided by a single ontology (e.g., stating that *Researcher* $\sqsubseteq$ *AcademicStaff*), or by reasoning over information spread in several ontologies (e.g., that *Researcher* $\sqsubseteq$ *ResearchStaff* in one ontology and that *ResearchStaff* $\sqsubseteq$ *AcademicStaff* in another). Experiments with Scarlet in the area of ontology matching and reported in [Sabou et al. 2008] show that this technique can lead to a high percentage (70%) of correct relations.

## 5. EVIDENCE INTEGRATION

This section discusses the data structures and methods used for the integration of the collected evidence. Table I exemplifies how the evidence sources ($es$) outlined in the previous section yield evidence to suggest candidate concepts, such as the one shown for the seed concept ($C_s$) *climate change*.

### 5.1 Semantic Network and Confidence Matrix

The *semantic network* is a data structure to represent all collected evidence. It connects collected terms with the seed ontology via directed weighted links. The *confidence matrix (CM)* is an additional and temporally extended form to store evidence data. It has the following characteristics:

—Each relation between a seed concept $C_s$ and a candidate concept $C_c$ is represented by its own CM.

—For every evidence source (extraction method), the CM contains the observed connection strength between the two concepts. Missing evidence for a source results in a value of 0.0.

—The second dimension is the temporal one. The presented system computes new ontologies in a particular domain at regular intervals. The temporal dimension tracks the history of connection strength in a fine-grained manner and thereby lays the foundation for ontology evolution experiments. The CM is also needed for adapting the source impact vectors (see below) according to user feedback.

Equation 1 displays the confidence matrix (CM) for the seed concept $C_s$ and candidate $C_c$ with the dimensions evidence source $es$ and time $t$ and the corresponding confidence values $c_{es_i,t_j}$.

$$CM_{C_s,C_c} = \begin{bmatrix} c_{es_1,t_1} & c_{es_1,t_2} & \cdots & c_{es_1,t_m} \\ c_{es_2,t_1} & c_{es_2,t_2} & \cdots & c_{es_2,t_m} \\ & & \vdots & \\ c_{es_n,t_1} & c_{es_n,t_2} & \cdots & c_{es_n,t_m} \end{bmatrix} \tag{1}$$

### 5.2 Concept Selection and Positioning

For the tasks of concept selection and positioning, the semantic network is transformed into a spreading activation network. Spreading activation is a search technique inspired by the cognitive models of the human brain; the processing involves iterations of sets of pulses sent through the network to find the most active nodes. Spreading activation manages the selection of a small number of new domain concepts from the large number of candidates and also supports positioning, i.e. the decision where to attach new concepts [Liu et al. 2005].

Our approach is novel in its use of an adaptive source impact vector (SIV) to influence spreading activation weights. Previous research used predefined source weights when transforming the semantic network to a spreading activation network. Those weights are multiplied with the observed connection strength – as stored in the CM – to determine the weighted links (i.e., the impact of a particular evidence source) for spreading activation. Equation 2 shows a SIV for a given point in time $t_i$ which contains the impact values $I$ for evidence sources $es_j$.

$$SIV_{t_i} = \begin{bmatrix} I_{es_1} & I_{es_2} & \cdots & I_{es_n} \end{bmatrix} \tag{2}$$

The source impact values adapt according to user feedback (see below), lowering the impact of sources which tend to yield low-quality evidence data and vice versa. Similarly to the confidence matrix, we trace the evolution of the SIV over time. The added temporal dimension transforms the SIV into a source impact matrix. We intialize the SIV (for time $t_0$) with preset values stemming from the existing system or metrics such as the Google PageRank.

### 5.3 Relation Detection

The concept positioning process described in the previous section not only results in a single relation between each newly added domain concept and a seed concept, but computes and outputs the relation strength from a new concept to all seed concepts – this data is valuable input for user evaluation and ontology evolution processes.

Relation detection deals with labeling relations of new concepts connected to seed concepts, but since it is not the focus of this work, we only briefly discuss it here. Relation detection involves the identification of taxonomic relations with techniques such as Hearst pattern, subsumption analysis and head noun analysis [Liu et al. 2005]. For the detection of non-taxonomic relations, vector space model similarities of verbs co-occurring with labeled and unlabeled relations suggest a label from a pre-defined set of relation types. To refine the results, the method combines those suggestions with ontology reasoning on external sources such as DBpedia and OpenCyc and finally applies semantic validation [Weichselbraun et al. 2010b].

## 6. USER FEEDBACK AND EVALUATION

This section covers two essential aspects of the presented approach: (i) collecting human evaluation feedback about the learned ontologies by means of games with a purpose (Section 6.1), which allows (ii) the automatic optimization of source impact values as part of the learning algorithm (Section 6.2). Human feedback and subsequent evaluation of the newly added ontology elements involves a three-step process performed after every ontology extension stage:

(1) *Relevance of the concept:* Evaluate the relevance of added concepts for the domain ontology. A straightforward classification schema (e.g. "very relevant", "slightly relevant", "slightly irrelevant", "not relevant") allows assessing the concepts. Based on the outcome of the user evaluation, concepts below the relevance threshold are pruned. In previous work, this task was conducted manually by domain experts.

(2) *Relevance of the relation:* Assess whether the seed concept and the new concept are related to each other (e.g., "fuel" and "fossil fuel"). If the relation between the seed and new concept is not relevant, then the user is presented with a list of alternative relations for the new concept. As discussed in Section 5.2, the spreading activation algorithm generates a relation strength value for any connection between seed and new concepts, which is used to select the most likely candidates for user evaluation.

(3) *Relevance of the relation type:* Determine whether suggested relation types such as "isA" and "causes" are correct, and suggest more appropriate relation types if required. Feedback on relation types yields a ranking of relevant relation types.

### 6.1 Human Computation

The term Human Computation (HC) describes systems that combine computers with large numbers of humans working together to complete a task that cannot readily be done by either alone, typically involving an automated system that serves up task instances to humans and then aggregates the results produced [Quinn and Bederson 2011]. HC tasks usually have an artificial intelligence flavour, when humans perform sub-tasks for which effective AI algorithms are not yet available. HC includes a number of genres, which can be distinguished along various dimensions, such as the motivation of human contributors (e.g. fun vs. altruism vs. payment) and the skills required of them, how individual results are aggregated and how quality is managed [Quinn and Bederson 2011]. Key HC genres include crowdsourcing, mechanised labour (e.g. Amazon vs. Mechanical Turk), and games with a purpose [Ahn and Dabbish 2008], where human contributors are motivated by formulating tasks as enjoyable games.

We use GWAPs as a modality for soliciting evaluations and player feedback - developing specific games for each of the tasks described above. These games can be seen as an extension of the authors' existing GWAP portfolio, which includes the *Sentiment Quiz* [Rafelsberger and Scharl 2009] for acquiring language resources needed to support sentiment detection, and the *Climate Quiz* depicted in Figure 4.

Crowdsourcing systems must address a number of challenges including (i) how to attract and retain players, (ii) how to ensure high quality data and (iii) how to aggregate results. Games made available via social networking sites typically attract more players than those published on stand-alone web sites. Additionally, these games can take advantage of viral advertising mechanisms, where players are encouraged to (and can easily) invite their friends to play the game. For quality control purposes, the GWAP approach relies on measuring inter-player agreement and we will also include questions that assess the users' domain expertise. Data aggregation can take various forms, ranging from numeric

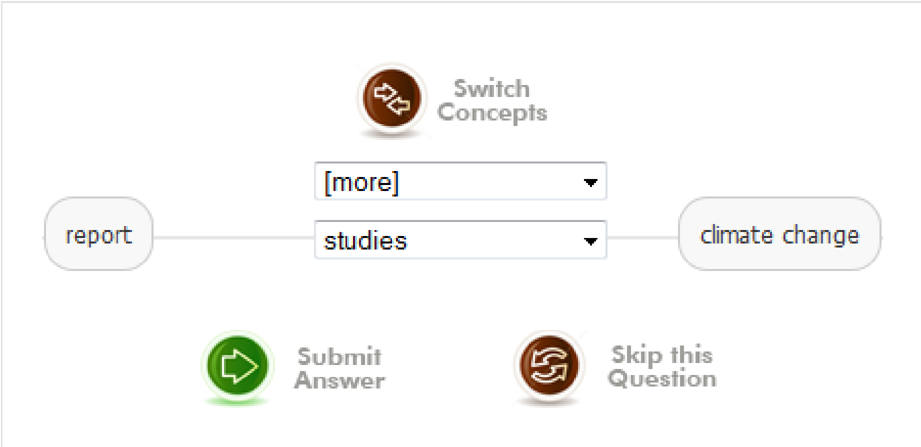


Fig. 4. A Facebook-based GWAP for eliciting semantic relations between terms.

integration to aggregating diverse user ratings. Conflicting user-generated content is integrated by means of the feedback component outlined in the next section.

### 6.2 Feedback Loops for Dynamic Ontology Learning

Section 5.2 introduces the source impact vector and its role in the concept selection and positioning algorithms. The feedback component (i) integrates potentially conflicting and heterogeneous user feedback, and (ii) adapts the initial values of the source impact vector accordingly.

The feedback component integrates heterogeneous feedback by applying voting rules (such as majority votes) that consider a user's impact which is determined by his expertise and track record in the vote aggregation process. This process yields an optimized ontology that corresponds to the users' perceptions of the domain, and to a corresponding source impact vector $SIV_{t_i}$. Using neural network learning techniques such as backpropagation allows adjusting the weights of the spreading activation network and the corresponding values of the source impact vector accordingly.

[Weichselbraun et al. 2010a] demonstrate that additional evidence sources improve the outcome of ontology learning systems as compared to text-only baseline calculations([Liu et al. 2005]). Two evaluation metrics (pointwise mutual information and domain expert judgement) helped assess the impact of adding social evidence sources such as Twitter and Technorati, which typically contain the latest domain terminology and trends. Regarding the quality of the learned ontology fragments, both evaluation metrics show the clear benefits of adding social sources, with significance scores for a Welch two sample t-test (for the PMI) and a Wilcoxon rank sum test with continuity correction (for the discrete expert ranking) exceeding 99.9%.

The inclusion of external sources posses the danger of shifting the ontology's focus away from the domain represented in the textual data towards the included evidence sources. The optimization of the learning algorithm presented in this article considers the sources' relevance in regard to the learning task by adapting SIVs on fine-grained evidence sources with GWAPs. This process increases the relevance of the resulting ontology by keeping the source impact vector within a predefined interval to prevent over-fitting to specific domains.

## 7. CONCLUSION AND OUTLOOK

This article presents an ontology learning architecture to integrate multiple, heterogeneous and evolving evidence sources (unstructured, structured, social) into a consistent data model, and uses a human computation approach for evaluating ontologies and optimizing the underlying learning algorithms. Specific data structures such as the confidence matrix allow capturing ontology evolution as part of a continuous learning process in a fine-grained and efficient manner. Starting from a seed ontology, the system automatically gathers evidence from text, from online social media and from Semantic Web sources into a joint graph data structure. Spreading activation algorithms generate new concept candidates and support the concept positioning process. Games with a purpose are integrated at various points of the learning cycle, e.g. to prune concept candidates not relevant to the domain, or the evaluate and eventually refine the concept positioning and relation detection processes. The result is an adaptive ontology learning system, which tightly integrates GWAP-collected user feedback for optimizing the learning algorithms.

Future work will not only enrich the number of tasks supported by the presented crowdsourcing framework, but also conduct fine-grained ontology evolution experiments assessing (i) the characteristics of confidence matrices resulting from specific datasets, and (ii) the use of separate SIVs for concept selection and concept positioning, since evidence sources are likely to have different suitability for those two tasks. These experiments will help to fine-tune the system, improve the quality of resulting ontologies, and classify patterns of ontology evolution.

## REFERENCES


Ahn, L. and Dabbish, L. Designing Games with a Purpose. *Communications of the ACM* 51 (8): 58–67, 2008.

Angeletou, S., Sabou, M., Specia, L., and Motta, E. Bridging the gap between folksonomies and the semantic web: An experience report. In *ESWC2007 Workshop: Bridging the Gap between Semantic Web and Web*. Innsbruck, Austria, pp. 30–43, 2007.

Cimiano, P., Maedche, A., Staab, S., and Voelker, J. Ontology learning. In *Handbook on Ontologies*, S. Staab and D. Rudi Studer (Eds.). International Handbooks on Information Systems. pp. 245–267, 2009.

Correndo, G., Alani, H., and Salvadores, M. Social Support for Ontological Mediation and Data Integration. *International Journal of Virtual Communities and Social Networking* 1 (3): 19–34, 2009.

d'Aquin, M., Baldassarre, C., Gridinoc, L., Sabou, M., Angeletou, S., and Motta, E. Watson: Supporting Next Generation Semantic Web Applications. In *WWW/Internet conference*. Vila real, Portugal, 2007.

d'Aquin, M., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V., and Guidi, D. Towards a New Generation of Semantic Web Applications. *IEEE Intelligent Systems* 23 (3): 20 – 28, 2008.

Gruber, T. R. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies* 43 (5-6): 907–928, 1995.

Hearst, M. A. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics*. Nantes, France, pp. 539–545, 1992.

Hees, J., Roth-Berghofer, T., and Biedert, R. BetterRelations: Detailed Evaluation of a Game to Rate Linked Data Triples. In *First International Workshop on Orderig and Reasoning*. Koblenz, Germany, 2011.

Kuo, Y., Lee, J., Chiang, K., Wang, R., Shen, E., Chan, C., and Hsu, J. Community-based game design: experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. Paris, France, pp. 15–22, 2009.

Liu, W., Weichselbraun, A., Scharl, A., and Chang, E. Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management* 0 (1): 50–58, 2005.

Lopez, V., Nikolov, A., Sabou, M., Uren, V., and Motta, E. Scaling up Question-Answering to Linked Data. In *Proceedings of the Knowledge Engineering and Management by the Masses*. Lisbon, Portugal, pp. 193–210, 2010.

Markotschi, T. and Voelker, J. Guess What?! Human Intelligence for Mining Linked Data. In *Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data; International Conference on Knowledge Engineering and Knowledge Management*. Lisbon, Portugal, 2010.

Maynard, D. and Aswani, N. Bottom-up Evolution of Networked Ontologies from Metadata (NeOn Deliverable D1.5.4), 2010.

Maynard, D., Funk, A., and Peters, W. SPRAT: A Tool for Automatic Semantic Pattern-based Ontology Population. In *International Conference for Digital Libraries and the Semantic Web*. Trento, Italy, 2009.

Mika, P. Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics* 5 (1): 5–15, 2007.

Navigli, R., Velardi, P., and Faralli, S. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. Barcelona, Spain, pp. 1872–1877, 2011.

Popitsch, N. and Haslhofer, B. DSNotify: handling broken links in the web of data. In *Proceedings of the 19th International Conference on World Wide Web*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarti (Eds.). Raleigh, NC, USA, pp. 761–770, 2010.

Quinn, A. J. and Bederson, B. B. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the Annual Conference on Human Factors in Computing Systems*. Vancouver, BC, Canada, pp. 1403–1412, 2011.

Rafelsberger, W. and Scharl, A. Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*. Torino, Italy, pp. 193–198, 2009.

Sabou, M., d'Aquin, M., and Motta, E. Exploring the Semantic Web as Background Knowledge for Ontology Matching. *Journal on Data Semantics* vol. 11, pp. 156–190, 2008.

Sabou, M., Fernandez, M., and Motta, E. Evaluating Semantic Relations by Exploring Ontologies on the Semantic Web. In *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems*. Saabruecken, Germany, 2009.

Sabou, M., Gracia, J., Angeletou, S., d'Aquin, M., and Motta, E. Evaluating the Semantic Web: A Task-based Approach. In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*. Busan, South Korea, pp. 423–437, 2007.

Sanchez, D. and Moreno, A. Learning Non-taxonomic Relationships from Web Documents for a Domain Ontology Construction. *Data and Knowledge Engineering* 64 (3): 600–623, 2008.

Siorpaes, K. and Hepp, M. Games with a Purpose for the Semantic Web. *IEEE Intelligent Systems* 23 (3): 50–60, 2008a.

Siorpaes, K. and Hepp, M. OntoGame: Weaving the semantic web by online games. In *Proceedings of the 5th European Semantic Web Conference*, S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis (Eds.). Vol. 5021. Tenerife, Spain, pp. 751–766, 2008b.

Specia, L. and Motta, E. Integrating folksonomies with the semantic web. In *Proceedings of the European Semantic Web Conference*. Lecture Notes in Computer Science, vol. 4519. Innsbruck, Austria, pp. 624–639, 2007.

Umbrich, J., Hausenblas, M., Hogan, A., Polleres, A., and Decker, S. Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web*, C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas (Eds.). Vol. 628. Raleigh, NC, USA, 2010.

Waitelonis, J., Ludwig, N., Knuth, M., and Sack, H. WhoKnows? - Evaluating Linked Data Heuristics with a Quiz that cleans up DBpedia. *International Journal of Interactive Technology and Smart Education (ITSE)* 8 (3): 236–248, 2011.

Weichselbraun, A., Wohlgenannt, G., and Scharl, A. Augmenting lightweight domain ontologies with social evidence sources. In *Proceedings of the 9th International Workshop on Web Semantics*, A. M. Tjoa and R. R. Wagner (Eds.). Bilbao, Spain, pp. 193–197, 2010a.

Weichselbraun, A., Wohlgenannt, G., and Scharl, A. Refining non-taxonomic relation labels with external structured data to support ontology learning. *Data & Knowledge Engineering* 69 (8): 763–778, 2010b.

Weichselbraun, A., Wohlgenannt, G., and Scharl, A. Applying Optimal Stopping Theory to Improve the Performance of Ontology Refinement Methods. In *Proceedings of 4th Hawaii International Conference on System Sciences*. Hawaii, USA, 2011.

Wohlgenannt, G., Weichselbraun, A., and Scharl, A. Integrating Structural Data into Methods for Labeling Relations in Domain Ontologies. In *20th International Workshop on Database and Expert Systems Application; 8th International Workshop on Web Semantics*. Linz, Austria, pp. 94–98, 2009.

Wolf, L., Knuth, M., Osterhoff, J., and Sack, H. RISQ! Renowned Individuals Semantic Quiz: a Jeopardy like quiz game for ranking facts. In *Proceedings of the 7th International Conference on Semantic Systems*. Graz, Austria, 2011.

Zablith, F., d'Aquin, M., Sabou, M., and Motta, E. Using Ontological Contexts to Assess the Relevance of Statements in Ontology Evolution. In *Proc. of EKAW - Knowledge Engineering and Management by the Masses*, P. Cimiano and H. S. Pinto (Eds.). Lecture Notes in Computer Science, vol. 6317. Lisbon, Portugal, pp. 226–240, 2010.