A Web Mining Tool to Validate Previously Discovered Semantic Correspondences. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY : 1–

KnoE: A Web Mining Tool to Validate Previously Discovered Semantic Correspondences

Jorge Martinez-Gil, *Member, ACM*, and José F. Aldana-Montes

University of Málaga, Department of Computer Languages and Computing Sciences
Boulevard Louis Pasteur 35. Postal code: 29071. Málaga, Spain.

E-mail: jorgemar@acm.org; jfam@lcc.uma.es

Received —

Revised month day, year

Abstract The problem of matching schemas or ontologies consists of providing the corresponding entities in two or more models of this kind that belong to a same domain but have been developed separately. Nowadays there are a lot of techniques and tools for addressing this problem, however, the complex nature of the matching problem means existing solutions for real situations are not fully satisfactory. On the other hand, the Google Similarity Distance has appeared recently. Its purpose is to mine knowledge from the Web using the Google Search Engine in order to compare semantically text expressions. Our work consists of developing a software application for validating results discovered by schema and ontology matching tools by using the philosophy behind this distance. Moreover, we are interested in using not only Google, but other popular search engines using this similarity distance. The results have revealed three main facts: firstly, some web search engines can help us to validate semantic correspondences satisfactorily. Secondly there are significant differences among the web search engines, and thirdly the best results are obtained when using combinations of the web search engines that we have studied.

Keywords Databases, Database Integration, Data and Knowledge Engineering Tools and Applications

1 Introduction

The Semantic Web is a new paradigm for the Web in which the semantics of information is defined, making it possible for the Web to understand and satisfy the requests of people and machines wishing to use the web resources. Therefore, most authors consider it as a vision of the Web from the point of view of an universal medium for data, information, and knowledge exchange [1].

In relation to knowledge, the notion of ontology as a form of representing a particular universe of discourse or some part of it is very important. Schema and ontology matching is a key aspect in order that the knowledge exchange in this extension of the Web may be real [2]; it allows organizations to model their own knowledge without having to stick to a specific standard. In fact, there are two good reasons why most organizations are not interested in working with a standard for modeling their own knowledge: (a) it is very difficult or expensive for many organizations to reach an agreement about a common standard, and (b) these standards do not often fit to the specific needs of the all participants in the standardization process.

Although ontology matching is perhaps the most valuable way to solve the problems of heterogeneity between information systems and, there are a lot of techniques for matching ontologies very accurately, experience tells us that the complex nature of the problem to be solved makes it difficult for these techniques to operate satisfactorily for all

Terms alignment and matching are often confused. In this work, we will call matching the task of finding correspondences between knowledge models and alignment to the output of the matching task

kinds of data, in all domains and as all users expect. Moreover the heterogeneity and ambiguity of data descriptions makes it unlikely that optimal mappings for many pairs of entities will be considered as *best mappings* by any of the existing matching algorithms.

Our opinion is shared by other colleagues who have also experienced this problem. In this way, experience tells us that getting such function is far from being trivial. As we commented earlier, for example, "finding good similarity functions is, data-, context-, and sometimes even user-dependent, and needs to be reconsidered every time new data or a new task is inspected" or "dealing with natural language often leads to a significant error rate" [3]. Figure 1 shows an example of matching between two ontologies developed from two different perspectives. Matching is possible because they belong to a common domain that we could name "world of transport", however there is difficult to find a function in order to discover all possible correspondences.

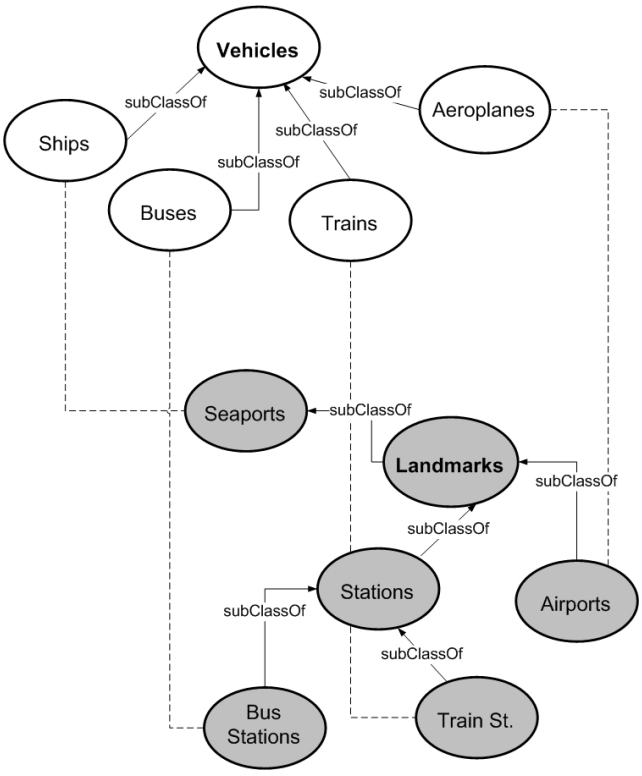

Fig. 1. Example of matching between two ontologies representing vehicles and landmarks respectively

As a result, new mechanisms have been developed from customized similarity measures [4, 5] to hybrid ontology matchers [6, 7], meta-matching systems [8, 9] or even soft computing techniques [10, 11]. However, results are still not entirely satisfactory, but we consider that the web knowledge could be the solution. Our idea is not entirely original; for example, web knowledge has already been used by Ernandes et al. [12] for solving crosswords automatically in the past.

We think that this a very promising research line. In fact, we are interested in three characteristics of the World Wide Web (WWW):

1. It is one of the biggest and most heterogeneous databases in the world. And possibly the most valuable source of general knowledge. Therefore, the Web fulfills the properties of Domain Independence, Universality and Maximum Coverage proposed by Gracia and Mena [13].

2. It is close to human language, and therefore can help to address problems related to natural language processing.

3. It provides mechanisms to separate relevant from non-relevant information or rather the search engines do so. We will use these search engines to our benefit.

In this way, we believe that the most outstanding contribution of this work is the foundation of a new technique which can help to identify the best web knowledge sources for solving the problem of validating semantic correspondences to match knowledge models satisfactorily. In fact, in [14], the authors state: "We present a new theory of similarity between words and phrases based on information distance and Kolmogorov complexity. To fix thoughts, we used the World Wide Web (WWW) as the database, and Google as the search engine. *The method is also applicable to other search engines* and databases". Our work is about those search engines.

Therefore in this work, we are going to mine the Web, using search engines to decide if a pair of semantic correspondences previously discovered by a schema or ontology matching tool could be

true. It should be taken into account that under no circumstances this work can be considered as a demonstration that one particular web search engine is better than another or that the information it provides is, in general, more accurate.

The rest of this article is organized as follows. Section 2 describes the problem statement related to the schema and ontology alignment problem and reviews some of the most outstanding matching approaches. Section 3 describes the preliminary definitions that are necessary for understanding our proposal. Section 4 deals with the details of KnoE, the tool we have built in order to test our hypothesis. Section 5 shows the empirical data that we have obtained from several experiments using the tool. Section 6 discusses the related works presented in the past, and finally, Section 7 describes the conclusions and future lines of research.

2 Problem Statement

The process of matching schemas and ontologies can be expressed as a function where given a couple of models of this kind, an optional input alignment, a set of configuration settings and a set of resources, a result is returned. The result returned by the function is called alignment. An alignment is a set of semantic correspondences (also called mappings) which are tuples consisting of a unique identifier of the correspondence, entities belonging to each of the respective ontologies, the type of correspondence (equality, generalization, specialization, etc..) between the entities and a real number between 0 and 1 representing the mathematical probability that the relationship described by R may be true. The entities that can be related are concepts, object properties, data properties, and even instances belonging to the models which are going to be matched.

According to the literature, we can group the subproblems related to schema and ontology matching in seven different categories.

1. How to obtain high quality alignments automatically.
2. How to obtain alignments in the shortest possible time.
3. How to identify the differences between matching strategies and determine how good each is according to the problem to be solved.
4. How to align very large models.
5. How to interact with the user during the process.
6. How to configure the parameters of the tools in an automatic and intelligent way.
7. How to explain to the user why this alignment was generated.

Most researchers work on some of these subproblems. Our work does not fit perfectly with any of them but it identifies a new one: How to validate previously discovered semantic correspondences. Therefore, we work with the output from existing matching tools (preferably with cutting-edge tools). There are a lot of outstanding approaches for implementing this kind of tools: [15, 16, 17, 18, 19, 20, 21]. They often use one or more of the following matching strategies:

1. **String normalization.** This consists of methods such as removing unnecessary words or symbols. Moreover, strings can be used for detecting plural nouns or to take into account common prefixes or suffixes as well as other natural language features.
2. **String similarity.** Text similarity is a string based method for identifying similar elements. For example, it may be used to identify identical concepts of two ontologies based on having a similar name [22].
3. **Data Type Comparison.** These methods compare the data type of the ontology elements. Similar concept attributes have to be of the same data type.
4. **Linguistic methods.** This consists of the inclusion of linguistic resources such as lexicons and thesauri to identify possible similarities. The most popular linguistic method is to use WordNet [23] to identify some kinds of relationships between entities.

5. **Inheritance analysis.** These kinds of methods take into account the inheritance between concepts to identify relationships. The most popular method is the analysis that tries to identify subsumptions between concepts.

6. **Data analysis.** These kinds of methods are based on the rule: If two concepts have the same instances, they will probably be similar. Sometimes, it is possible to identify the meaning of an upper level entity by looking at one of a lower level.

7. **Graph-Mapping.** This consists of identifying similar graph structures in two ontologies. These methods use known graph algorithms. Mostly this involves computing and comparing paths, children and taxonomy leaves [4].

8. **Statistical analysis.** This consists of extracting keywords and textual descriptions to detect the meaning of one entity in relation to others [24].

9. **Taxonomic analysis.** It tries to identify similar concepts or properties by looking at their related entities. The main idea behind this analysis is that two concepts belonging to different ontologies have a certain degree of probability of being identical if they have the same neighborhood [25].

10. **Semantic analysis.** According to [2], semantic algorithms handle the input based on its semantic interpretation. One supposes that if two entities are the same, then they share the same interpretations. Thus, they are deductive methods. Most outstanding approaches are propositional satisfiability and description logics reasoning techniques.

Most of these strategies have proved their effectiveness when they are used with some kind of synthetic benchmarks like the one offered by the Ontology Alignment Evaluation Initiative (OAEI) [26]. However, when they process real ontologies, their results are worse [27]. For this reason, we propose to use a kind of linguistic resources which have not been studied in depth in this field. Our approach consists of mining knowledge from the Web with the help of web search engines, in this way, we propose to get benefit from the fact that this kind of knowledge is able to support the process of validating the set of correspondences belonging to an schema or ontology alignment.

On the other hand, several authors have used web knowledge in their respective work, or have used a generalization: background knowledge [28, 29, 30, 31]. This uses all kinds of knowledge sources to extract information: dictionaries, thesauri, document collections, search engines and so on. For this reason web knowledge is often considered a more specific subtype.

The classical approach to this problem has been addressed in literature with the use of a tool called WordNet [23]. Related to this approach, the proposals presented in [15] is the most remarkable. The advantage that our proposal presents in relation to the use of WordNet [23] is that it reflects more closely the language used by people to create their content on the Internet, therefore, it is much closer to everyday terms, thus, if two words appear very often on the same website, we believe that there is some probability that a semantic relationship exists between them.

There are other works about Web Measures. For instance, Gracia and Mena [13] try to formalize a measure for comparing the relatedness of two terms using several search engines. Our work differs from that in several key points. Firstly, they use Yahoo! as a search engine in their experiment arguing its balance between good correlation with human judgment and fast response time. Instead we prefer to determine the best source by means of an empirical study. Secondly, authors say they can perform ontology matching tasks with their measure. Based in our experiences, this is not a great idea; i.e. they need to launch many thousands queries in a search engine in order to align two small ontologies and to lower the tolerance threshold [27]. Therefore, they obtain a lot of false positives. Instead, we propose to use the cutting-edge tool [21] to match schemas or ontologies and use

web knowledge to validate these previously discovered correspondences. For the same ontologies, we need a thousand times fewer queries number and we do not incur any additional false positive.

3 Technical Preliminaries

In this section, we are going to explain some technical details which are necessary to understand our proposal.

Definition 1 (Similarity measure). *A similarity measure sm is a function* $sm : \mu_1 \times \mu_2 \mapsto R$ *that associates the similarity between two entities* μ_1 *and* μ_2 *to a similarity score* $sc \in \Re$ *in the range [0, 1].*

A similarity score of 0 stands for complete inequality and 1 for equality of the entities μ_1 and μ_2.

Definition 2 (Alignment). *An alignment a is a set of tuples* $\{(id, e, e', n, R)\}$. *Where id is an identifier of the mapping, e and* e' *are entities belonging to two different models, R is the relation of correspondence between these entities, and n is a real number between 0 and 1 that represents the probability that R may be true.*

Definition 3 (Matching function). *A matching function mf is a function* $mf : O_1 \times O_2 \overset{sm}{\rightarrow} A$ *that associates two input knowledge models* km_1 *and* km_2 *to an alignment a using a similarity measure.*

There are many matching techniques for implementing this kind of function as we shown in Section II.

Definition 4 (Alignment Evaluation). *An alignment evaluation ae is a function* $ae : a \times a_R \mapsto precision \in \Re \in [0, 1] \times recall \in \Re \in [0, 1]$ *that associates an alignment a and a reference alignment* a_R *to two real numbers stating the precision, recall of a in relation to* a_R.

Precision states the faction of retrieved correspondences that are relevant for a matching task. Recall is the fraction of the relevant mappings that are obtained successfully in a matching task. In this way, precision is a measure of exactness and recall a measure of completeness. The problem here is that techniques can be optimized either to obtain high precision at the cost of the recall or, alternatively, recall can be optimized at the cost of the precision. For this reason a measure, called f-measure, is defined as a weighting factor between precision and recall. For the rest of this work, we use the most common configuration which consists of weighting precision and recall equally.

Definition 5 (Relatedness Distance). *Relatedness Distance is a metric function that states how related two or more entities belonging to different models are and meets the following axioms*

1. *relatedness(a, b)* $\leq$ *1*
2. *relatedness(a, b) = 1 if and only if a = b*
3. *relatedness(a, b) = relatedness(b, a)*
4. *relatedness(a, c)* $\leq$ *relatedness(a, b) + relatedness(b, c)*

Notions of similarity and relatedness seems to be very similar, but they are not. Similarity expresses equivalence, while relatedness expresses membership in a common domain of discourse. For example, similarity between *car* and *wheel* is low while they are not equivalent at all, while relatedness between *car* and *wheel* is high. We can express the differences more formally:

Theorem 1 (Similarity involves relatedness). *Let* μ_1 *and* μ_2 *be two entities belonging to different knowledge models. If* μ_1 *and* μ_2 *are similar then* μ_1 *and* μ_2 *are related.*

Theorem 2 (Relatedness does not involve similarity). *Let* μ_1 *and* μ_2 *be two related entities belonging to different knowledge models. If* μ_1 *and* μ_2 *are related then we cannot guarantee that they are similar.*

Lemma 1 (About the validation of semantic correspondences). *Let S be the set of semantic correspondences generated using a specific technique. If any of these correspondences are not related, then they are false positives.*

Example 1 (About Lemma 1). *Let (bucks, bank, =, 0.8) be a mapping automatically detected by a matching tool. If we use a relatedness distance which, for example, tell us that bucks and bank do not co-occur in the same websites frequently, then we have that the matching tool generated a false positive. Otherwise, if bucks and bank co-occur very often in the Web, then we cannot refute the correctness of this mapping.*

Definition 6 (Hit). *Hit is an item found by a search engine to match specified search conditions. More formally, we can define a hit as the function $hit : \vartheta \mapsto N$ which associates a natural number to a set of words to ascertain its popularity in the WWW.*

A value of 0 stands for no popularity and the bigger the value, the bigger its associated popularity. Moreover, we want to remark that the function *hit* has many possible implementations. In fact, every web search engine implements it a different way. For this reason, we can not take into account only one search engine to perform our work.

Example 2. (Normalized Google Distance). *It is a measure of relatedness derived from the number of hits returned by the Google search engine for a given (set of) keyword(s). Keywords with the same or similar meanings in a natural language sense tend to be close in units of Google distance, while words with dissimilar meanings tend to be farther apart.*

The normalized Google distance (NGD) between two search terms a and a is

$$D(a,b) = \frac{mx\{\log hit(a), \log hit(b)\} - \log hit(a,b)}{\log M - mn\{\log hit(a), \log hit(b)\}} \quad (1)$$

where M is the total number of web pages searched by Google; $hit(a)$ and $hit(b)$ are the number of hits for search terms a and b, respectively; and $hit(a,b)$ is the number of web pages on which a and b co-occur.

Finally, we define a correspondence validator as an software artifact that uses a relatedness distance to detect false positives in schema or ontology alignments according to the Lemma 1. We have built a correspondence validator called Knowledge Extractor (KnoE).

4 KnoE

Semantic similarity between text expressions changes over time and across domains. The traditional approach to solve this problem has consisted of using manually compiled taxonomies. The problem is that a lot of terms are not covered by dictionaries; therefore, similarity measures that are based on dictionaries cannot be used directly in these tasks. However, we think that the great advances in web research have provided new opportunities for developing new solutions.

In fact, with the increase of larger and larger collections of data resources on the WWW, the study of web measures has become one of the most active areas for researchers. We consider that techniques of this kind are very useful for solving problems related to semantic similarity because new expressions are constantly being created and also new senses are assigned to existing expressions.

The philosophy behind KnoE (Knowledge Extractor) is to use a web measure based on the Google Similarity Distance [14]. This similarity measure gives us an idea of the number of times that two concepts appear together in comparison with the number of times that the two concepts appear separately in the subset from the Web indexed by a given search engine.

For the implementation of the function *hit*, we have chosen the following search engines from among the most popular in the ranking Alexa [32]: Google, Yahoo!, Lycos, Altavista, MSN and Ask.

The comparison is made between previously discovered correspondences. In this way we can decide if compared correspondences are considered reliable, or if they are not.

We could launch a task to make a comparison between all the entities of source and target knowledge model respectively. Then, only pairs of entities likely to be true (those whose parameter n exceeds a certain threshold) would be included in the final output alignment. There are several reasons why we do not propose this: Attempting to match models using directly such web knowledge function as Google Distance would involve considerable cost in terms of time and broadband consumption because each comparison needs 3 queries for the search engine and repeating this $m \cdot n$ times, where m and n are the number of entities belonging to the source and target knowledge models respectively. But the most important reason is that the amount of generated false positives means that this process may be unworkable. We have tried to solve the benchmark from OAEI [26] using only web knowledge and have obtained an average f-measure of about 19 percent. This represents a very low figure if we consider that the most outstanding tools obtains a f-measure of above 90 percent for the same benchmark [27].

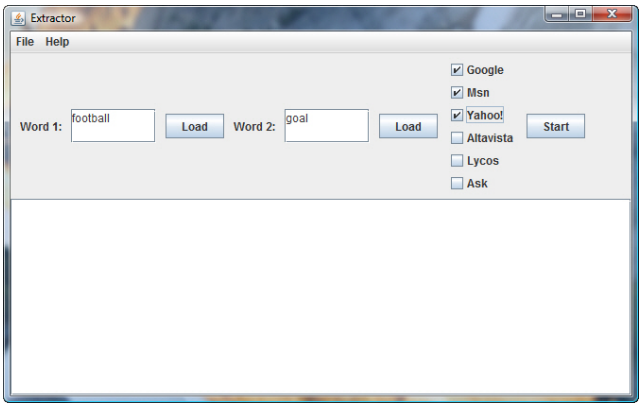

Fig. 2. Screenshot from the main window of KnoE. Users can select individual terms or lists. Moreover, they can choose some search engines for mining the web

Finally, KnoE has been coded using Java so it can be used in console mode on several operating systems, but to make the tool more friendly to the user, we have programmed a graphical user interface, as Figure 2 shows.

The operation mode is simple: once users select correspondences to compare, they should choose one or more search engines to perform the validation. In Figure 3, we have launched a task to validate the correspondence $(football, goal)$ using Google, Yahoo! and MSN. As it can be seen, Google considers that is not possible to refute the correctness of the correspondence, while Yahoo! and MSN consider that the equivalence is wrong.

Fig. 3. Graphical User Interface for KnoE. In this figure we show the validation of the pair *(football, goal)* according to several search engines

5 Empirical Evaluation

Now we evaluate KnoE using three widely accepted benchmark datasets. These benchmarks are Miller-Charles [33], Gracia-Mena [13], and Rubenstein-Goodenough [34] which are pairs of terms that vary from low to high semantic relatedness.

Several notes that are important in order to perform these experiments are: Some of the companies which own the web search engines do not allow many queries to be launched daily, because it is considered as mining service. So the service is limited and several days were necessary to perform the experiments. Results from Lycos Search Engine have not been included because, after several executions, they do not seem to be appropriate. In addition, it is important to note that this experiment was performed in February 2010, because the information indexed by the web search engines is

not static.

Table 1 shows the results that we have obtained for the Miller-Charles benchmark dataset. Table 2 shows the results we have obtained for the Gracia-Mena benchmark dataset. Finally, Table 3 shows the results we have obtained for the Rubenstein-Goodenough benchmark dataset.

On the other hand, Figures 4, 5, and 6 show the behavior of the average means from the web search engines in relation to the benchmark datasets. We have chosen to represent the average mean because it gives us the best result among the statistical functions studied. We have studied the mode and median additionally, but it does not outperform the average mean.

The comparison between the benchmark datasets and our results is made using the Pearson's Correlation Coefficient, which is an statistical measure which allows to compare two matrices of numeric values. Therefore the results can be in the interval [-1, 1], where -1 represents the worst case (totally different values) and 1 represents the best case (totally equivalent values).

- Experimental results on Miller-Charles benchmark dataset show that the proposed measure outperforms all the existing web-based semantic similarity measures by a wide margin, achieving a correlation coefficient of 0.61.

- Experimental results on Gracia-Mena benchmark dataset show that the proposed measure outperforms all the existing web-based semantic similarity measures (except Ask), achieving a correlation coefficient of 0.70.

- Experimental results on Rubenstein-Goodenough benchmark dataset show that the proposed measure outperforms all the existing web-based semantic similarity measures (except Yahoo!) achieving a correlation coefficient of 0.51.

The average mean presents a better behavior than the rest of studied mining processes: It is the best for the first benchmark dataset and the second one for the second and third benchmark dataset. We interpret this in the following form: although a correct pair of concepts cannot be validated by a specific search engine, it is very difficult that all search engines can be wrong at the same time. Therefore, for the rest of this work, we are going to use the average mean in our semantic correspondence validation processes.

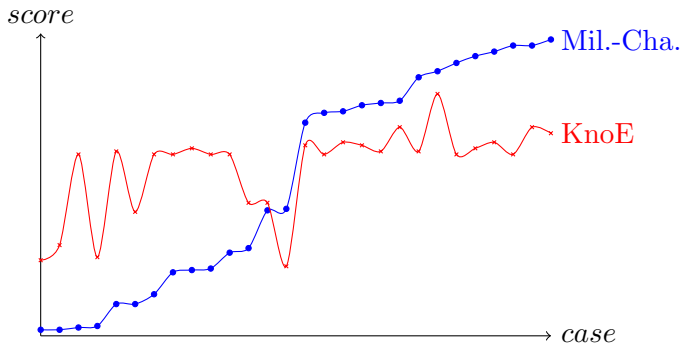

Fig. 4. Graphic representation of the behavior for the Miller-Charles benchmark

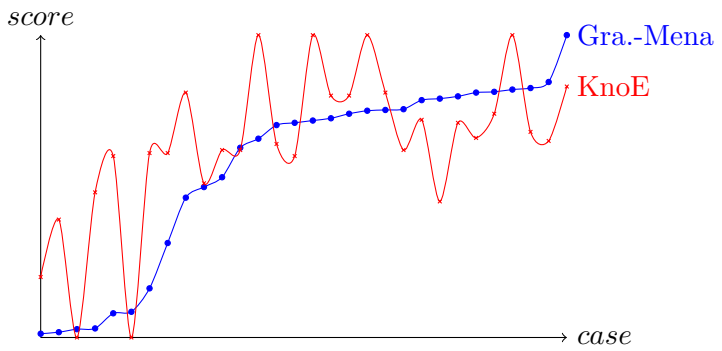

Fig. 5. Graphic representation of the behavior for the Gracia-Mena benchmark

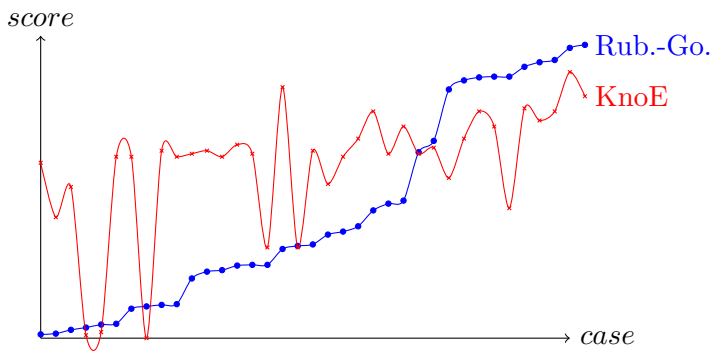

Fig. 6. Graphic representation of the behavior for the Rubenstein-Goodenough benchmark

	Mil.-Cha.	**Google**	**Ask**	**Altavista**	**MSN**	**Yahoo**	**KnoE**
cord-smile	0.13	0.05	0.25	1.00	0.00	0.00	0.26
rooster-voyage	0.08	0.24	1.00	0.00	0.00	0.00	0.25
noon-string	0.08	0.50	1.00	0.00	0.00	0.00	0.30
glass-magician	0.11	1.00	1.00	1.00	0.00	0.01	0.60
monk-slave	0.55	1.00	1.00	1.00	0.00	0.00	0.60
coast-forest	0.42	1.00	1.00	1.00	0.02	0.01	0.61
monk-oracle	1.10	1.00	1.00	1.00	0.00	0.00	0.60
lad-wizard	0.42	0.04	1.00	1.00	0.00	0.00	0.41
forest-graveyard	0.84	1.00	1.00	1.00	0.00	0.01	0.60
food-rooster	0.89	1.00	1.00	1.00	0.00	0.00	0.60
coast-hill	0.87	1.00	1.00	1.00	0.06	0.02	0.62
car-journey	1.16	0.17	1.00	1.00	0.01	0.00	0.44
crane-implement	1.68	0.14	1.00	0.00	0.00	0.00	0.23
brother-lad	1.66	0.18	1.00	1.00	0.00	0.00	0.44
bird-crane	2.97	1.00	1.00	1.00	0.13	0.09	0.64
bird-cock	3.05	1.00	1.00	1.00	0.07	0.07	0.63
food-fruit	3.08	1.00	1.00	1.00	0.03	0.01	0.61
brother-monk	2.82	1.00	1.00	1.00	0.11	0.04	0.63
asylum-madhouse	3.61	1.00	1.00	1.00	0.00	0.00	0.60
furnace-stove	3.11	0.46	1.00	1.00	0.00	1.00	0.69
magician-wizard	3.50	1.00	1.00	1.00	0.04	0.98	0.80
journey-voyage	3.84	1.00	1.00	1.00	0.00	0.00	0.60
coast-shore	3.70	1.00	1.00	1.00	0.02	0.08	0.62
implement-tool	2.95	1.00	1.00	1.00	0.00	0.02	0.60
boy-lad	3.76	1.00	1.00	1.00	0.18	0.02	0.64
automobile-car	3.92	1.00	1.00	1.00	0.01	0.34	0.67
midday-noon	3.42	1.00	1.00	1.00	0.07	0.00	0.61
gem-jewel	3.84	1.00	1.00	1.00	0.39	0.05	0.69
Correlation	1.00	0.47	0.26	0.35	0.43	0.34	**0.61**

Table 1. Experimental results obtained on Miller-Charles benchmark dataset

	Gra.-Mena	**Google**	**Ask**	**Altavista**	**MSN**	**Yahoo**	**KnoE**
transfusion-guitar	0.05	0.02	0.00	1.00	0.00	0.00	0.20
xenon-soul	0.07	0.58	0.38	1.00	0.01	0.00	0.39
nanometer-feeling	0.11	0.00	0.00	0.00	0.00	0.00	0.00
blood-keyboard	0.12	1.00	0.41	1.00	0.00	0.01	0.48
cloud-computer	0.32	1.00	1.00	1.00	0.00	0.01	0.60
theorem-wife	0.34	0.00	0.00	0.00	0.00	0.00	0.00
pen-lamp	0.65	1.00	1.00	1.00	0.06	0.00	0.61
power-healing	1.25	1.00	1.00	1.00	0.04	0.03	0.61
city-river	1.85	1.00	1.00	1.00	0.53	0.52	0.81
theft-house	1.99	0.55	1.00	1.00	0.01	0.00	0.51
professional-actor	2.12	1.00	1.00	1.00	0.02	0.10	0.62
dog-friend	2.51	1.00	1.00	1.00	0.08	0.01	0.62
atom-bomb	2.63	1.00	1.00	1.00	1.00	1.00	1.00
computer-calculator	2.81	1.00	1.00	1.00	0.01	0.20	0.64
person-soul	2.84	1.00	1.00	1.00	0.01	0.00	0.60
sea-salt	2.87	1.00	1.00	1.00	1.00	1.00	1.00
pencil-paper	2.90	1.00	1.00	1.00	0.07	0.93	0.80
penguin-Antarctica	2.96	1.00	1.00	1.00	0.00	1.00	0.80
yes-no	3.00	1.00	1.00	1.00	1.00	1.00	1.00
ten-twelve	3.01	1.00	1.00	1.00	1.00	0.06	0.81
car-wheel	3.02	1.00	1.00	1.00	0.11	0.01	0.62
car-driver	3.14	1.00	1.00	1.00	0.12	0.46	0.72
letter-message	3.16	0.24	1.00	1.00	0.01	0.01	0.45
river-lake	3.19	1.00	1.00	1.00	0.29	0.27	0.71
citizen-city	3.24	1.00	1.00	1.00	0.02	0.28	0.66
keyboard-computer	3.25	1.00	1.00	1.00	0.03	0.67	0.74
blood-transfusion	3.28	1.00	1.00	1.00	1.00	1.00	1.00
mathematics-theorem	3.30	1.00	1.00	1.00	0.00	0.39	0.68
hour-minute	3.38	1.00	1.00	1.00	0.16	0.08	0.65
person-person	4.00	1.00	1.00	1.00	0.15	1.00	0.83
Total	1.00	0.54	**0.74**	0.44	0.32	0.54	0.70

Table 2. Experimental results obtained on Gracia-Mena benchmark dataset

	Rub.-Goo.	**Google**	**Ask**	**Altavista**	**MSN**	**Yahoo**	**KnoE**
fruit-furnace	0.05	0.88	1.00	1.00	0.00	0.00	0.58
autograph-shore	0.06	1.00	1.00	0.00	0.00	0.00	0.40
automobile-wizard	0.11	0.51	1.00	1.00	0.00	0.00	0.50
mound-stove	0.14	0.00	0.00	0.00	0.00	0.00	0.00
grin-implement	0.18	0.00	0.00	0.00	0.00	0.00	0.00
asylum-fruit	0.19	1.00	1.00	1.00	0.01	0.00	0.60
asylum-monk	0.39	1.00	1.00	1.00	0.00	0.00	0.60
graveyard-madhouse	0.42	0.00	0.00	0.00	0.00	0.00	0.00
boy-rooster	0.44	1.00	1.00	1.00	0.05	0.03	0.62
cushion-jewel	0.45	1.00	1.00	1.00	0.00	0.00	0.60
asylum-cemetery	0.79	1.00	1.00	1.00	0.00	0.07	0.61
grin-lad	0.88	1.00	1.00	1.00	0.11	0.01	0.62
shore-woodland	0.90	1.00	1.00	1.00	0.02	0.00	0.60
boy-sage	0.96	1.00	1.00	1.00	0.17	0.01	0.64
automobile-cushion	0.97	1.00	1.00	1.00	0.00	0.06	0.61
mound-shore	0.97	0.52	1.00	0.00	0.00	0.00	0.30
cemetery-woodland	1.18	1.00	1.00	1.00	1.00	0.15	0.83
...	...	...	...	...	...	...	...
crane-rooster	1.41	1.00	1.00	1.00	0.00	0.00	0.60
hill-woodland	1.48	1.00	1.00	1.00	0.08	0.22	0.66
cemetery-mound	1.69	0.65	1.00	1.00	0.10	1.00	0.75
glass-jewel	1.78	1.00	1.00	1.00	0.01	0.02	0.61
magician-oracle	1.82	1.00	1.00	1.00	0.49	0.00	0.70
sage-wizard	2.46	1.00	1.00	1.00	0.02	0.01	0.61
oracle-sage	2.61	1.00	1.00	1.00	0.18	0.00	0.63
hill-mound	3.29	0.38	1.00	1.00	0.10	0.15	0.53
cord-string	3.41	1.00	1.00	1.00	0.11	0.21	0.66
glass-tumbler	3.45	1.00	1.00	1.00	0.36	0.41	0.75
grin-smile	3.46	1.00	1.00	1.00	0.34	0.14	0.70
serf-slave	3.46	0.15	1.00	1.00	0.00	0.00	0.43
autograph-signature	3.59	1.00	1.00	1.00	0.20	0.59	0.76
forest-woodland	3.65	1.00	1.00	1.00	0.03	0.58	0.72
cock-rooster	3.68	1.00	1.00	1.00	0.20	0.56	0.75
cushion-pillow	3.84	1.00	1.00	1.00	0.40	1.00	0.88
cemetery-graveyard	3.88	1.00	1.00	1.00	0.00	1.00	0.8
Correlation	1.00	0.21	0.33	0.39	0.29	**0.63**	0.51

Table 3. Experimental results obtained on Rubenstein-Goodenough benchmark dataset

5.1 Correspondence Validation

There are two kinds of correspondences in an alignment: correct mappings and false positives. Correct mappings are correspondences between entities belonging to two different models which are true. False positives are correspondences between entities belonging to two different models which are false, but the technique that generated the alignment considered them as true. To reduce the number of false positives in a given alignment increases the recall and, therefore, improves the quality of the alignment.

On the other hand, our strategy can face four different situations: to validate or not validate correct mappings and to validate or not validate false positives. Obviously, we want that correct mappings may be validated and that false positives may be not validated. Under no circumstances we want correct mappings to not be validated; that means not only that we are not improving the results, but we are getting worse them (by decreasing the precision). Validated false positives neither improve nor diminish the precision or recall, for this reason it is a failure, although it does not alter the overall quality of the results.

In Table 4 we can see a sample for real results obtained when validating an alignment between two ontologies related to bibliography. We have chosen a threshold of 0.51, thus, all correspondences with a relatedness score higher than this value will be validated. There are a total count of 18 discovered semantic correspondences. 6 false positives has not been validated, so we have improved the recall a 33 percent. 2 correct mappings have not been validated, so the precision has decreased a 11 percent. Finally, a false positive has not been validated, so the quality has not been altered. With this results (recall increased 33 percent and precision decreased 11 percent) the overall quality of the alignment (f-measure) has been improved a 11 percent using KnoE.

5.2 Discussion

The results we have obtained can give us an idea of the behavior of different web search engines and their possible application to validate our strategy for schema and ontology matching. In fact, we can highlight two features which draw attention in the set of results that we have obtained:

1. There is great disparity between the results obtained by the web search engines that have been taken into account. We think it would be especially interesting to know why.

2. The average of the values from the different search engines outperform, in general, the values returned by the web search engines atomically.

Regarding the first fact, we must look at how search engines treat the identical words, synonyms and word variations. We can see many cases with totally opposite results. This shows that there are web knowledge sources that are more appropriate than others, at least, for the domain in which the study has been performed.

Why such search engines like Yahoo! offer better results than other search engines, we are not sure. At first, we could consider that it is either the quantity or quality of content indexed by these search engines. On the other hand, Ask indexes currently much less web contents than either Google or Yahoo!, but the treatment of queries and/or indexing of content that is relevant to the datasets used, means that it can also provide good results according to some of these benchmarks. In this way, we think that the results that we have obtained do not depend largely on the indexed content.

Secondly, we have that the average mean of the single results is, in general, better that the web search engines. We have obtained good results for the average mean in the three cases, 0.61, 0.7 and 0.51 respectively. These results are on average better that the rest of the single web measures, what means that this configuration could be useful to validate semantic correspondences.

6 Related Works

Apart from semantic correspondence validation, web measures can be used in many other applications [13], such as analysis of texts, annotation of resources, information retrieval, automatic

Mapping	**Type**	**Relatedness**	**Action**	**Status**
articles-papers	Correct Mapping	0.61	Validated	Hit
book-booklet	False Positive	0.41	Not Validated	Hit
book-bookPart	False Positive	0.10	Not Validated	Hit
city-town	Correct Mapping	0.59	Validated	Hit
chapters-sections	Correct Mapping	0.60	Validated	Hit
communications-talks	Correct Mapping	0.60	Validated	Hit
conference-congress	Correct Mapping	0.61	Validated	Hit
Institution-Organization	Correct Mapping	0.61	Validated	Hit
name-FirstName	False Positive	0.45	Not Validated	Hit
parts-tomes	Correct Mapping	0.13	Not Validated	Failure
person-PersonList	False Positive	0.45	Not Validated	Hit
periodicity-frequency	Correct Mapping	0.47	Not Validated	Failure
publisher-published	False Positive	0.60	Validated	Failure
size-dimensions	Correct Mapping	0.80	Validated	Hit
TechReport-report	False Positive	0.40	Not Validated	Hit
unpublished-manuscript	Correct Mapping	1.00	Validated	Hit
unpublished-publisher	False Positive	0.40	Not Validated	Hit
url-link	Correct Mapping	0.86	Validated	Hit

Table 4. Sample from the results obtained when validating a real alignment. A threshold of 0.51 has been defined empirically

indexing, or spelling correction, as well as entity resolution. On the other hand, we have identified three points when researching about web measures: the Web as a knowledge corpus, measures based on web hits, and measures based on text snippets.

- Regarding the Web as a knowledge corpus has become an active research topic recently. For instance, that unsupervised models demonstrably perform better when n-gram counts are obtained from the Web rather than from other corpus was presented in [35]. Resnik and Smith [36] extracted sentences from the Web to create a parallel corpora for machine translation.

- Regarding web hits, Turney [37] defined a point-wise mutual information measure using the number of hits returned by a Web search engine to recognize synonyms. Matsuo et al. [38] proposed the use of Web hits for extracting communities on the Web. They measured the association between two personal names using the overlap coefficient, which is calculated based on the number of Web hits for each individual name and their conjunction.

- There is other way to measure relatedness: text snippets from search engines. For example, Sahami et al. [39] measured semantic similarity between two queries using snippets returned for those queries by a search engine. For each query, they collect snippets from a search engine and represent each snippet as a TF-IDF-weighted term vector. Chen et al. [40] proposed a double-checking model using text snippets returned by a Web search engine to compute semantic similarity between words.

To the best of our knowledge, our proposal is the first attempt to use web knowledge in order to improve the schema and matching process by supervising alignments automatically, so unfortunately, we do not have references to compare quantitatively with our work yet.

7 Conclusions and Future Work

In this work we have presented a proposal for validating previously discovered semantic correspondences using web knowledge. It has mainly consisted of developing the concepts of relatedness and building a tool for implementing the Google Similarity Distance [11] using other popular search engines. With the obtained results we can assign to the discovered semantic correspondences a degree of confidence, and therefore we can discard or include them in the final alignment which will be presented to the users. In this way, we are able to improve the recall, and therefore, the overall quality (f-measure) of the results. In our work we can extract that:

1. Web Search Engines can be considered valid sources of knowledge that provide support to the task of validated semantic correspondences in a completely unsupervised manner.

2. There is a wide disparity in the results generated by the web search engines that we have studied.

3. Ask, Google, and Yahoo! seem to be the best web knowledge sources for validating previously semantic correspondences. However, an average mean of all search engines is, in general, even better. We think that these results are no dependent on a greater quantity of content indexed and higher quality, however, the treatment they gives to user queries makes them the most appropriate web search engines to perform this task from among the benchmark that we have studied.

As future work, we propose a comparison of the knowledge provided by WordNet [23] and that provided by the web sources. On the other hand, the development of a new version for the tool to automate the entire process, from the selection of discovered correspondences to determining the best knowledge sources to validate them, is already in process. The idea is to be able to evaluate web sources automatically according to widely accepted benchmarks. Our end goal is, given the specifications of an ontology matching problem, to

compute the optimum alignment function so that the problem can be solved accurately and without requiring human intervention in any part of the process. In this way, the semantic interoperability between people, computers or simply agents might become true.

Acknowledgements

We wish to thank to the anonymous reviewers for the comments and suggestions which have helped to improve this work. We thank to Lisa Huckfield for proofreading this manuscript. This work has been funded by Spanish Ministry of Innovation and Science through REALIDAD: Gestion, Analisis y Explotacion Eficiente de Datos Vinculados (TIN2011-25840).

References

[1] Berners-Lee T, Hendler J, Lassila O. The semantic web. *Scientific American*, 2001, 284(5):34–43.

[2] Euzenat J, Shvaiko P. Ontology matching. Springer, 2007.

[3] Kiefer C, Bernstein A, Stocker M. The fundamentals of isparql: A virtual triple approach for similarity-based semantic web tasks. ISWC/ASWC, 2007, pp.295–309.

[4] Ziegler P, Kiefer C, Sturm C, Dittrich K R, Bernstein A. Detecting similarities in ontologies with the soqa-simpack toolkit. EDBT, 2006, pp.59–76.

[5] Lambrix P, Tan H. A tool for evaluating ontology alignment strategies. *J. Data Semantics*, 2007, 8:182–202.

[6] Domshlak C, Gal A, Roitman H. Rank aggregation for automatic schema matching. *IEEE Trans. Knowl. Data Eng.*, 2007, 19(4):538–553.

[7] Gal A, Anaby-Tavor A, Trombetta A, Montesi D. A framework for modeling and evaluating automatic semantic reconciliation. *VLDB J.*, 2005, 14(1):50–67.

[8] Ehrig M, Staab S, Sure Y. Bootstrapping ontology alignment methods with apfel. International Semantic Web Conference, 2005, pp.186–200.

[9] Lee Y, Sayyadian M, Doan A, Rosenthal A. etuner: tuning schema matching software using synthetic scenarios. *VLDB J.*, 2007, 16(1):97–122.

[10] Mao M, Peng Y, Spring M. An adaptive ontology mapping approach with neural network based constraint satisfaction. *J. Web Sem.*, 2010, 8(1):14–25.

[11] Wang J, Ding Z, Jiang C. Gaom: Genetic algorithm based ontology matching. APSCC, 2006, pp.617–620.

[12] Ernandes M, Angelini G, Gori M. Webcrow: A web-based system for crossword solving. AAAI, 2005, pp.1412–1417.

[13] Gracia J, Mena E. Web-based measure of semantic relatedness. WISE, 2008, pp.136–150.

[14] Cilibrasi R, Vitányi P M B. The google similarity distance. *IEEE Trans. Knowl. Data Eng.*, 2007, 19(3):370–383.

[15] Budanitsky A, Hirst G. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 2006, 32(1):13–47.

[16] Motta E, Sabou M. Next generation semantic web applications. ASWC, 2006, pp.24–29.

[17] Do H H, Rahm E. Coma - a system for flexible combination of schema matching approaches. VLDB, 2002, pp.610–621.

[18] Aumueller D, Do H H, Massmann S, Rahm E. Schema and ontology matching with coma++. SIGMOD Conference, 2005, pp.906–908.

[19] Drumm C, Schmitt M, Do H H, Rahm E. Quickmig: automatic schema matching for data migration projects. CIKM, 2007, pp.107–116.

[20] Ehrig M, Sure Y. Foam - framework for ontology alignment and mapping - results of the ontology alignment evaluation initiative. Integrating Ontologies, 2005.

[21] Wang Z, Zhang X, Hou L, Zhao Y, Li J, Qi Y, Tang J. Rimom results for oaei 2010. OM, 2010.

[22] Navarro G. A guided tour to approximate string matching. *ACM Comput. Surv.*, 2001, 33(1):31–88.

[23] Miller G A. Wordnet: A lexical database for english. *Commun. ACM*, 1995, 38(11):39–41.

[24] Gil J M, Montes J F A. Evaluation of two heuristic approaches to solve the ontology meta-matching problem. *Knowl. Inf. Syst.*, 2011, 26(2):225–247.

[25] Avesani P, Giunchiglia F, Yatskevich M. A large scale taxonomy mapping evaluation. International Semantic Web Conference, 2005, pp.67–81.

[26] Euzenat J, Meilicke C, Stuckenschmidt H, Shvaiko P, dos Santos C T. Ontology alignment evaluation initiative: Six years of experience. *J. Data Semantics*, 2011, 15:158–192.

[27] Shvaiko P, Euzenat J, Giunchiglia F, He B (eds.). *Proceedings of the 2nd International Workshop on Ontology Matching (OM-2007) Collocated with the 6th International Semantic Web Conference (ISWC-2007) and the 2nd Asian Semantic Web Conference (ASWC-2007), Busan, Korea, November 11, 2007*, *CEUR Workshop Proceedings 304*, CEUR-WS.org, 2008.

[28] van Harmelen F. Two obvious intuitions: Ontology-mapping needs background knowledge and approximation. IAT, 2006, p.11.

[29] Giunchiglia F, Shvaiko P, Yatskevich M. Discovering missing background knowledge in ontology matching. ECAI, 2006, pp.382–386.

[30] Vazquez R, Swoboda N. Combining the semantic web with the web as background knowledge for ontology mapping. OTM Conferences (1), 2007, pp.814–831.

[31] Gligorov R, ten Kate W, Aleksovski Z, van Harmelen F. Using google distance to weight approximate ontology matches. WWW, 2007, pp.767–776.

[32] Ranking A. 2009. Http://www.alexa.com.

[33] Miller G A, Charles W G. Contextual correlates of semantic similarity. *Language Cognitive Processes*, 1991, 6(1):1–28.

[34] Rubenstein H, Goodenough J B. Contextual correlates of synonymy. *Commun. ACM*, 1965, 8(10):627–633.

[35] Keller F, Lapata M. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 2003, 29(3):459–484.

[36] Resnik P, Smith N A. The web as a parallel corpus. *Computational Linguistics*, 2003, 29(3):349–380.

[37] Turney P D. Mining the web for synonyms: Pmi-ir versus lsa on toefl. *CoRR*, 2002, cs.LG/0212033.

[38] Matsuo Y, Sakaki T, Uchiyama K, Ishizuka M. Graph-based word clustering using a web search engine. EMNLP, 2006, pp.542–550.

[39] Sahami M, Heilman T D. A web-based kernel function for measuring the similarity of short text snippets. WWW, 2006, pp.377–386.

[40] Chen H H, Lin M S, Wei Y C. Novel association measures using web search with double checking. ACL, 2006.

Jorge Martinez-Gil received his PhD from University of Malaga in 2010. He currently holds an acting professorship within the University of Extremadura (Spain). Dr. Martinez-Gil has published more than 30 scientific papers, including some published by prestigious journals.

José F. Aldana-Montes is currently a professor from the University of Malaga (Spain) and Head of Khaos Research. Dr Aldana-Montes has more than 20 years of experience in research about several aspects of databases, semistructured data and semantic technologies.