O-032

A search engine for semantic web

Mamdouh Farouk, Mitsura Ishizuka

Creative Informatics Department, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 1138656, Japan
mamdouh@mi.ci.i.u-tokyo.ac.jp , ishizuka@i.u-tokyo.ac.jp

1. Abstract

Consuming semantically represented data in semantic web is an important for the next generation of the web. This paper presents a new search engine, which has important features. This search engine depends on Concept Description Language (CDL). CDL format enables web agents to deep understand web content. Moreover, it does not depend on ontology. It is a statement-based search, which can answer query semantically. It uses inference to get more information depending on a set of available related rules. The result of this search engine is more relative to users' needs.

2. Introduction

Semantic search engine seeks for a more accurate query results. It is not like normal search engine that is based on keyword matching. Many query expansion are utilized to improve keyword searching [1]. However, semantic search is based on meaning matching. The first important step towards semantic search is the representation format for data sets. In other words, the data sets to be searched should be represented into a machine understandable format. Consequently, the search engine can understand the data set. The semantic search enables normal web users to get the benefits of semantic web [2]. Semantic search engine make the use of semantic web to improve the web search [3].

Resource Description Framework RDF is the most popular format that is used to represent web resources. RDF is w3c recommendation and is widely used by researcher in semantic web field. There are many RDF ontologies in different domains available on the Internet. These ontologies can be used to describe web resources into RDF format.

In this paper, we propose a search engine that is based on Concept Description Language (CDL). It also use user defined rules to answer more queries.

This paper is organized as follows: section 2 explain CDL language and why we are using. Section 3 descript the proposed approach in details. Implementation and experiments are discussed in section 4. Section 5 concludes this paper and shows the future work.

3. Concept Description Language (CDL)

Our approach uses a new semantic language which called Concept Description Language (CDL). This semantic format, which proposed by Institute of Semantic Computing, describes semantic/conceptual structure of contents (resources) and can deal with natural languages, mathematical expressions, movie, music, etc [4]. The aims of CDL are to realize machine understandability of web text contents, and to overcome language barrier on the web [5].

CDL is one of the three forms that can be used to express CWL (Common Web Language). Moreover, CWL is a part of the Incubator Activity of W3C [6]. CWL is a common language for exchanging information through the web and also for enabling computers to process information semantically.

This new representation bases on Concept Description Language for natural language (CDL.nl) which describes the concept structure of the text based on a set of predefined semantic relations [4]. The main advantage of CDL is that it does not depend on ontologies. However it depends on the Universal Networking Language Knowledge Based (UNLKB) and a set of universal relations so it can be used without facing similar problems of using ontologies. CDL depends on a complete dictionary which contains around 120,000 words. Moreover, this dictionary contains the definitions of these words represented into CDL form. This means that the computer can understand the word definition as well as the semantic relation between words that is also contained in that dictionary. Consequently, reasoning agents can use these definitions to get better understanding and more accurate results. For example, the representation of a statement such as "John bought a computer yesterday" in CDL is: *{#A Event tmp='past';*
{#a1 buy(agt>person,obj>thing);}
{#a2 computer(icl>machine);}
{#a3 yesterday(icl>day);}   {John John;}
[#a1 agt John]   [#a1 obj #a2]   [#a1 tim #a3]}

In the above notation the first three lines represent concepts used in the statement and the last line represents the relations between these concepts.

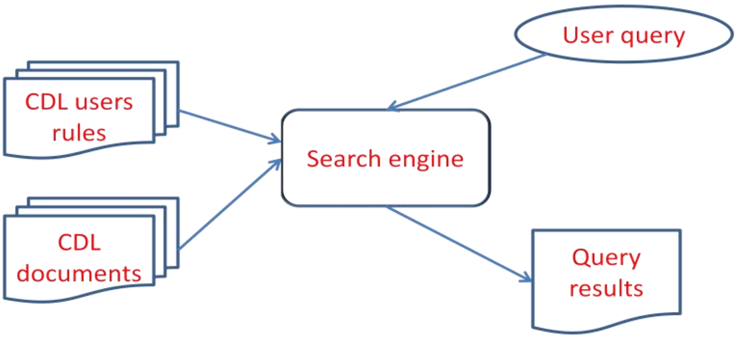


Figure 1: general architecture of the proposed search

4. The proposed System

The proposed system takes a CDL query as input and returns the answer of this query, fig.1. CDL Query Builder tool can be used to facilitate creating CDL query. The query engine searches the available CDL document to find the query answer. The search engine tries to matching query statement with the CDL document statements.

4.1 The semantic data

As mentioned before this approach uses CDL format. Some data already available on the Internet in CDL format, www.tuics.com. It is not a large data set. We use this data to show the visibility of our approach. This data are generated using our previous work [7]. It is important to notice that there is no ambiguity in CDL data because each concept in CDL statement contains its usage form. In other words, when a concept is used in a CDL statement the user should select the concept form that is suitable to what he/she means.

4.2 Semantic Search

Our CDL semantic search technique is based on inexact graph matching. Each CDL statement is considered as a graph in which the nodes represent concepts and arcs represent relations. The proposed search does inexact matching between query graph and CDL statements graph. The query graph contains a variable node that the user asks about. By matching two graphs, this variable will be initiated with the proper value.

Matching graphs In this matching process, we use depth first technique with backtracking to match graphs because of the simplicity of the graphs to be matched.

Matching concepts The concept in CDL format contains two parts the word and the description. For example the following CDL concept, take(icl>get(agt>person,obj>thing)), has the word take and the description (icl>get(agt>person,obj>thing)). In concept matching, we try to make corresponding parts together. So, in case of complete matching, the two corresponding parts should be matched. The value of matching is 1 in such case. However, if there is one part is not matched it will be partially matched. Furthermore, the description part can be match using the CDL Dictionary hierarchy. The name part can be also match using synonyms or related words.

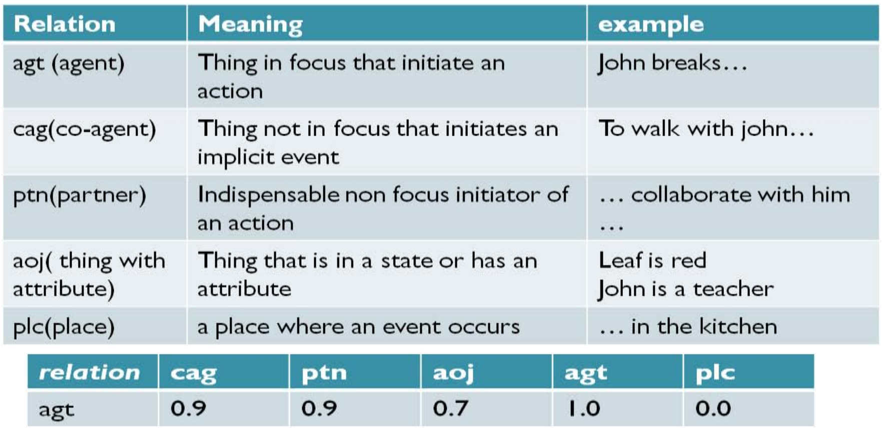
Matching variables There are two forms of variables that the user can use. The first form is the absolute form and the second is constrained form. Absolute form like (?K) and the constrained form like (?K(iof>person)). Therefore, the constrained variable is a variable with a specific description. The constrained variable can match only with the concepts that its description matches description of the variable. However, the absolute variable can match any concept or variable.

Matching relations The relation in CDL compounds from relation name, from concept and to concept. Consequently, Matching relation process contains two sub-routine matching relation name and matching corresponding concepts.

Relation name matching The propose approach does not match the names of the relation based on the keyword. However, matching the relation names is based on the relation meaning. Fig.2 shows the similarity values between different CDL relation. These similarity values is manually assigned based on the meaning of the relation and how it similar to other relations.

The second sub-routine is corresponding concept matching. In this sub-routine we just invoke the concept matching module.

The user can used the CDL query builder tool, Fig.3, to create CDL query accurately and easily. The user should enter a list of concepts and then he/she should enter the relation between these concepts. CDL Query Builder facilitates choosing usage form of the concept by showing different usage from of the entered concept based on CDL dictionary.



Relation	Meaning	example
agt (agent)	Thing in focus that initiate an action	John breaks…
cag(co-agent)	Thing not in focus that initiates an implicit event	To walk with john…
ptn(partner)	Indispensable non focus initiator of an action	… collaborate with him …
aoj(thing with attribute)	Thing that is in a state or has an attribute	Leaf is red John is a teacher
plc(place)	a place where an event occurs	… in the kitchen

relation	cag	ptn	aoj	agt	plc
agt	0.9	0.9	0.7	1.0	0.0

Figure 2. CDL relations similarity values

One of the important advantages of the proposed approach is that it uses a set of user defined rules to infer additional information to be used in query answering. The proposed search engine can make inference to get additional information based on a set of user defined rules. Backward chaining is used to fire the rules. Therefore, in case of matching between query and an action part of a rule, a new query is created from the condition part of the rule.

5. Experiments and Results.

A prototype of the proposed approach is implemented using java. The user should provide the developed tool by both CDL documents and CDL rules. After creating the CDL query the user should press search button. The query result appears in the bottom part of the tool. The result id a CDL statement that matches the user query. The result statement may be not in the CDL documents. However, it may be inferred using the inference engine. The following is an Example created using the developed tool.

Query Example

```
<q>
  <text>where does Ali work?</text>
  <UWs>
  <uw code="23">Ali(iof>person)</uw>
  <uw code="4D">work(agt>person)</uw>
  <uw code="9A">?L</uw>
  </UWs>
<relations>
  <r name="agt" from="4D" to="23" />
  <r name="plc" from="4D" to="9A" />
  </relations>
  </q>
```

Result:

23 Ali(iof>person)

5r work(agt>person)

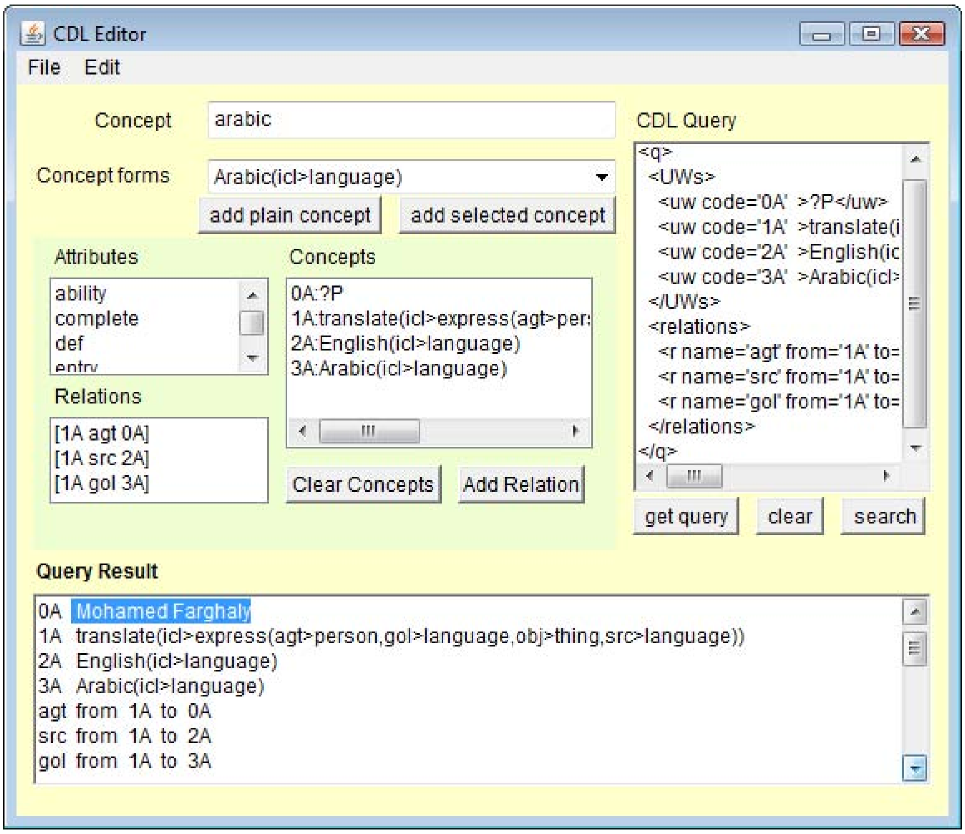
3a: Tokyo University

agt from="**5r**" to="**23**"

plc from="**5r**" to="**3a**”

6. Conclusion and future work

As a conclusion, the proposed approach for semantic search is depend on CDL language. Using CDL overcomes the problems of using different domain ontologies. The proposed approach also can infer additional information based on a user defined set of rules. The implemented prototype shows the visibility of the proposed approach and the search result is relevant to users’ needs.



Concept: arabic

Concept forms: Arabic(icl>language)

Attributes:
```
ability
complete
def
entry
```

Concepts:
```
0A:?P
1A:translate(icl>express(agt>per
2A:English(icl>language)
3A:Arabic(icl>language)
```

Relations:
```
[1A agt 0A]
[1A src 2A]
[1A gol 3A]
```

CDL Query:
```
<q>
  <UWs>
    <uw code='0A' >?P</uw>
    <uw code='1A' >translate(i
    <uw code='2A' >English(ic
    <uw code='3A' >Arabic(icl>
  </UWs>
  <relations>
    <r name='agt' from='1A' to=
    <r name='src' from='1A' to=
    <r name='gol' from='1A' to=
  </relations>
</q>
```

Query Result:
```
0A  Mohamed Farghaly
1A  translate(icl>express(agt>person,gol>language,obj>thing,src>language))
2A  English(icl>language)
3A  Arabic(icl>language)
agt  from  1A  to  0A
src  from  1A  to  2A
gol  from  1A  to  3A
```

Figure 3. CDL Query Builder tool

As a future work, we should complete the engine implementation. More experiments on a large data sets are needed to show the effectiveness of the proposed approach. Implement the engine as a web based search engine is also important.

7. References

[1] M¨akel¨a, E.: Survey of semantic search research. In: Proceedings of the Seminar on Knowledge Management on the Semantic Web, Department of Computer Science, University of Helsinki (2005)

[2] Lei, Y., Uren, V., Motta, E.: Semsearch: A search engine for the semantic web. In: Staab, S.,Svátek, V. (eds.) EKAW2006. LNCS (LNAI), vol. 4248, pp. 238–245. Springer, Heidelberg (2006)

[3] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, and J. Sachs. Swoogle: A semantic web search and metadata engine. In Proc. 13th ACM Conf.on Information and Knowledge Management, Nov. 2004.

[4] T. Yokoi, H. Uchida, K. Hasida, el al.CDL (Concept Description Language): A Common Language for Semantic Computing, www2005 workshop on the semantic computing initiative (SeC2005)

[5] Mitsuru Ishizuka, "A Common Concept Description of Natural Language Texts as the Foundation of Semantic Computing on the Web", IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, Taiwan, June 2008, p.385

[6] H. Uchida, T. Yokoi, M. Zhu, N. Saito, V. Avetisyan, “Common Web Language”, W3C Incubator Group Report, http://www.w3.org/2005/Incubator/cwl/XGR-cwl/, March 2008

[7] Mamdouh Farouk, Mitsuru Ishizuka. Semantic Structure Content for Dynamic Web Pages. In Proceedings of Web *Intelligence/IAT Workshops'2010. pp.253~256*