Alternative Approaches for Ontology Matching

M.Gawich
French University
Shorouk,Cairo
Egypt

A.Badr
Cairo University
Guiza
Egypt

H.Ismael
Modern University
Maadi,Cairo
Egypt

A.hegazy
Arab Academy
Sheraton,Cairo
Egypt

ABSTRACT

Ontology matching is generally defined as the process of finding correspondences between entities of different ontologies. It can help the data integration between autonomous agents, web services composition, and P2P information sharing. This process is applied through the use of ontology matching tools which use one or more ontology matching techniques. This paper presents tools which have been published in this field, such as Prompt [7], Smatch [5] and Ontobuilder [16]. Moreover the paper illustrates the drawbacks of these tools. New two tools are proposed to handle these drawbacks. The new proposed and other tools are tested using GlycO [8]and EnzyO[10] in the biochemistry field, Osteoarthritis and Rheumatoid in the medical field.

General Terms

Ontology matching techniques, Ontology matching tools.

Keywords

Ontology Matching, Ontology Alignment, Ontology Prune Ontology, Semantic Web.

1. INTRODUCTION

The development of communications and information technologies has led to the emergence of huge amount of heterogeneous information which needs semantic technologies to manage it. One of these technologies is the ontology matching.

Ontology is defined as formal explicit specification of shared conceptualization[2]. Where " formal" means that the ontology should be machine readable and "shared" means that it is validated by a group or community. Ontology[3] is applied in many dynamic applications which require ultimately matching operation; e.g.: agents, web service and peer to peer systems. First for the agents, they are autonomous software entities. Agent system should find the answer to a question the user asks for, one of agent system component which helps the agents to find the reliable answer is the knowledge base which can be represented by the ontology. Agents have to communicate each other in order to find the suitable answer; the communication is executed by the "FIPA" language. During this communication, the two agents find a difficulty to understand each other if they don't share the same ontology. The ontology matching is used to solve this problem. Second for the semantic web services, which can be described with regard to ontologies, a central common ontology can be imposed for finding the adequate services and for interfacing services. Ontology matching plays the role of a bridge between different ontologies. For example: if two organizations deal with dictionaries program , one is an electronic website its goal is to sell the dictionary program as a product . The second one is a library which enables the user to download a free dictionary program. The common product between them is the dictionary; the seller of the electronic commerce website is concerned with the version and dictionary price, its version, the compatibility with the different operating systems. The second is concerned with the dictionary type, its language, its price and its version. Both of them are concerned with the version. Finally for the peer to peer information sharing, the autonomous peer to peer use different terminologies and metadata models in order to represent their data , even if they share data in the same domain of interest thus , in order to establish meaningful information exchange between peers , the matching between their ontologies is rendered.

Ontology matching 'ontology mapping' is the process of exploring the similarity between two or more heterogeneous ontologies. Matching operation takes ontology files as input and determines the correspondences between the entities of the two ontologies. These correspondences are called "Alignment".

Figure [Fig.1] below shows the general matching process which is considered as a function which receives ontology file1 "O", ontology file 2 "O′ " as two ontology files and returns "A" as the alignment between the two ontologies.

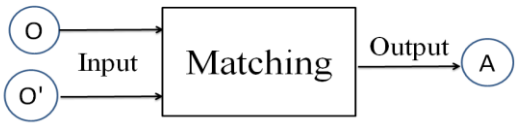

Fig 1: Ontology Matching Process.

The correspondence can either be expressed by one-to-one function or one-to-many function. One-to-one function denotes an entity in ontology which can only have one similar entity in another one, whereas one-to-many function addresses the fact that an entity may have more than one similar entity in another ontology [1].

Smatch, Prompt and Ontobuilder have some drawbacks. Never the less, Smatch has its very salient drawback. It uses WordNet, but can't detect the different entities which have the same meaning. Also, Prompt has its own drawbacks. In spite of the fact, that its output reflects any identical entities that have the same structure and the use of lexical matching with synonym algorithm, entities that have the same meaning aren't detected. Whereas the Ontobuilder isn't examined in

this paper, as it allows the user to automatically create ontology from the website address which contradicts the concept of ontology which should be validated by a group or community.

This paper presents in section (2) ontology matching techniques. Section (3) investigates the related work which shows a survey on the ontology matching tools. Section (4) introduces a proposal for two new ontology matching tools which apply two new ontology matching techniques. Finally section(5) presents a framework for comparative analysis between the current tools and the two proposed tools.

2. ONTOLOGY MATCHING TECHNIQUES

Ontology matching aims to find relations between entities "classes" expressed in different ontologies. These relations are equivalence relations that are discovered by the use of similarity measure between the entities of ontologies. Ontology matching techniques [4] are classified as Element, Extensional and semantic level techniques.

2.1 Element Level Techniques

Deal with ontology entities and instances in isolation with their relations. It contains the following techniques:

2.1.1 String-Based Techniques

Focus on the string structure, string based methods will find classes arthritis and Osteoarthritis are similar strings, but can't detect the similarity between Osteoarthritis and stiffness.
The major methods that are used: normalization, string equality technique, similarity technique, edit instance technique, statistical measures and path comparison technique.

2.1.2 Language -Based Methods

In language phenomenon, strings become text. Texts can be segmented to words easily identified sequence of letters that are derived from an entry in a dictionary. Language-based methods rely on using Natural Language Processing (NLP) techniques to help extract the meaningful terms from a text. They are classified to:

2.1.2.1 Intrinsic Methods

They involve tokenization, lemmatization, stop word elimination and term extraction.

2.1.2.2 Extrinsic Methods

They use external resources like dictionaries, lexicon, terminologies and Thesauri.

2.2 Extensional Techniques

These techniques focus on instance 'sub entities or subclasses'. They are suitable to apply the matching between two ontologies that share the same set of individuals "instances. Extensional methods are divided in three categories.

2.2.1 Common Extension Comparison

It is used to test the instances intersection between classes. Suppose A and B are similar classes. When A $\cap$ B =A =B, more general A $\cap$ B =A, A $\cap$ B =B,it involves the use of hamming distance and jaccard similarity techniques.

2.2.2 Disjoint extension comparison

It implies methods that can be based on statistical measures which are derived from the features of class members. Example: Statistical approach, Similarity based extension comparison and Matching-based comparison.

2.3 Semantic-Based Techniques

2.3.1 Techniques Based on External ontologies:

This section focuses on using intermediate formal ontologies. Intermediate ontologies can define the common context or background knowledge for the two ontologies to be matched.

The common ground can often be found by relating the ontologies to external resources which are classified to:

2.3.1.1 Breadth

They are general purpose resources or domain specific resources.

2.3.1.2 Formality

It implies the use of pure ontologies with semantic descriptions or informal resources, such as WordNet, or the use of formal resource such as DOLCE or the formal model of anatomy.

2.3.1.3 Status

It implies the use of resources which are considered as references, such as ontologies, thesauri or sets of instances.

2.3.2 Deductive Techniques:

It is used for testing the satisfiability of alignments. They are classified to propositional satisfiability, modal satisfiability techniques or description logic based techniques.

2.3.2.1 Propositional Satisfiability :

It involves the building of axioms "theory", a matching formula for each pair of classes c and c' from two ontologies and checks the validation of the formula.

2.3.2.2 Description Logic Techniques:

They imply the subsumption test that can be used to establish the relations between classes in a purely semantic manner. In fact, first merging two ontologies (after renaming) and then testing each pair of concepts and roles for subsumption is enough for matching terms with the same interpretation.

3. RELATED WORK

In this section, the dataset ontology files which will be used as Input are explored. Ontology matching tools are also presented.

3.1 Input

Ontology matching tools are tested by four ontology files, the first two files relates to the biochemistry field, while the two others relate to the medical field.

3.1.1 Biochemistry Input

3.1.1.1 GlycO

It explores [8] Glycans classification its reactions and chemical entities, also it explores the relationship between glycans and its chemical entities. The file format is '.owl'. Glycans are defined as complex carbohydrate structures, their role in the human body is the maintenance and the development of living cells. "Glycans[9] are built from simpler monosaccharide residues e.g mannose and glucose.These residues build the nodes of tree structure that are composed of chemical entities links between other residues . The synthesis of these glycans in organisms is an intricate process that can be modeled as a collection of biosynthetic pathways. At each step in such a pathway, an

enzyme-catalyzed reaction 'adds' a new residue as a leaf to an existing structure or 'moves' a whole subtree to a different parent."

3.1.1.2 EnzyO

Enzyme activity [10] plays a important role in the synthesis of glycans. They are subset of proteins. "Enzyme ontology [11] EnzyO keeps track of enzymes that catalyze the actions which produces the glycan structures. The ontology keeps track of basic information about enzymes for example their enzyme commission number (EC) , their protein structure as well as associations with genes that codes for it and the reactions it participates in".

3.1.2 Medical Input:

3.1.2.1 Rheumatoid

"Rheumatoid arthritis[12](RA) is an inflammatory disease that exerts its greatest impact on thosejoints of the body that are lined with synovium, a specialised tissue responsible for maintainingthe nutrition and lubrication of the joint. The distribution of joints affected (synovial joints) is characteristic." The Rheumatoid usually attacks human who suffers from surplus of immunity.

3.1.2.2 Ostreoarthritis

"Osteoarthritis [10] is the most common form of arthritis. It causes pain, swelling and reduced motion in your joints. It can occur in any joint, but usually it affects your hands, knees, hips or spine. Osteoarthritis breaks down the cartilage in the joints. Cartilage is the slippery tissue that covers the ends of bones in a joint. Healthy cartilage absorbs the shock of movement."

Rheumatoid and Osteoarthritis ontologies are built informally through the help of an expert in the immunology domain; Rheumatoid and Osteoarthritis ontologies are built formally by the protégé tool [15]. Both of them have common symptoms and share a common drug to eliminate the symptoms.

3.2 Ontology Matching Tools

3.2.1 Smatch

The Semantic matching (Smatch) [5] is a tool which applies a type of ontology matching techniques. It relies on semantic information encoded as xml ,the encoded information can be can be database or ontology . the tool enables the user to create or import two ontologies files encoded as xml nodes as figure[Fig.2] shows; the Smatch matcher identifies these nodes in two structures which semantically correspond to another one. For example, if the user imports the two ontologies files "c.xml" and "w.xml" which locate in "test data" folder in the Smatch tool, the Smatch matcher can detect that folder(class) :"College of Engineering" of the ontology file "c.xml" corresponds to "Materials Sciences and Engineering" folder "class" of the ontology file"w.xml" as figure[Fig.3] shows. Smatch can identify this correspondence because they are synonyms in English. This information is taken from a linguistic resource like WordNet.

WordNet [6] is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

The Smatch accepts only the ontology file with xml format as input. The ontology files must have xml format and include the xml tags as nodes. The Smatch applies the string based technique and language based methods. Concerning the output, Smatch detects the similarity "equivalence" between entities only if they have the same superclass , and shows other relations "disjoint,more general relations".

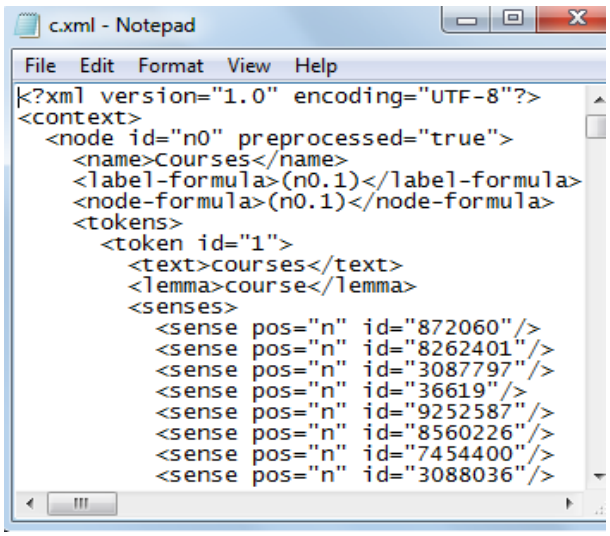

Fig 2: Smatch xml file

3.2.2 Prompt

The Prompt [7] is a plug-in suite for Protégé used to manage multiple ontologies. It has five main functions:

- Compare between an existing ontology and a different version of the same ontology.
- Map between two ontologies and browsing the common data.
- Move frames between the current including project and one of the included projects;
- Merge between two ontologies and add the resulting merged ontology to the current project;
- Extract a portion of ontology and add it to the current project.

Map function is the most relevant to the subject of this paper. The map function enables the match of the current ontology with another one based on lexical matching algorithm, or lexical matching algorithm with synonyms.

Prompt Protégé accepts the ontology file with the following format owl, xml and rdf-xml. It applies the string-based technique. It shows the similar entities of the two ontology files.

The following figure [Fig.4] shows the creation of synonym property for Smoking_cessation class. The value of its synonym is "Smoking_limitation". Although the use of lexical matching with synonym, entities which have the same meaning are not detected as figure[Fig.5] shows.

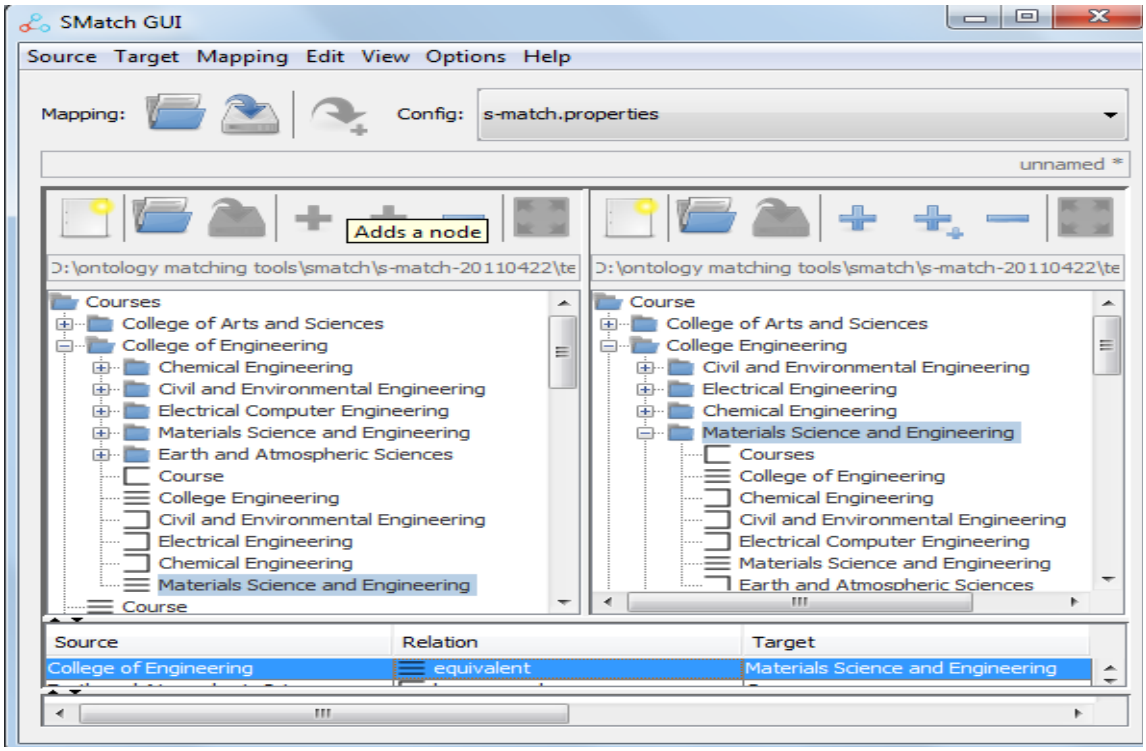

Fig 3 : Ontology Matching In Smatch

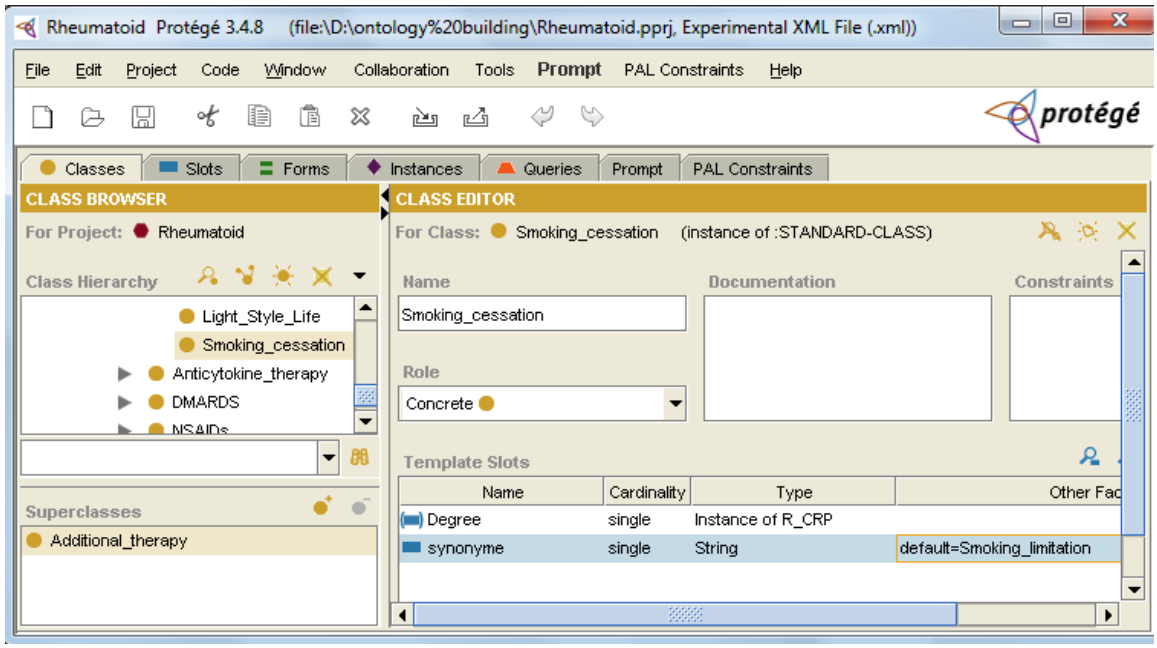

Fig 4 : Synonym Property Creation

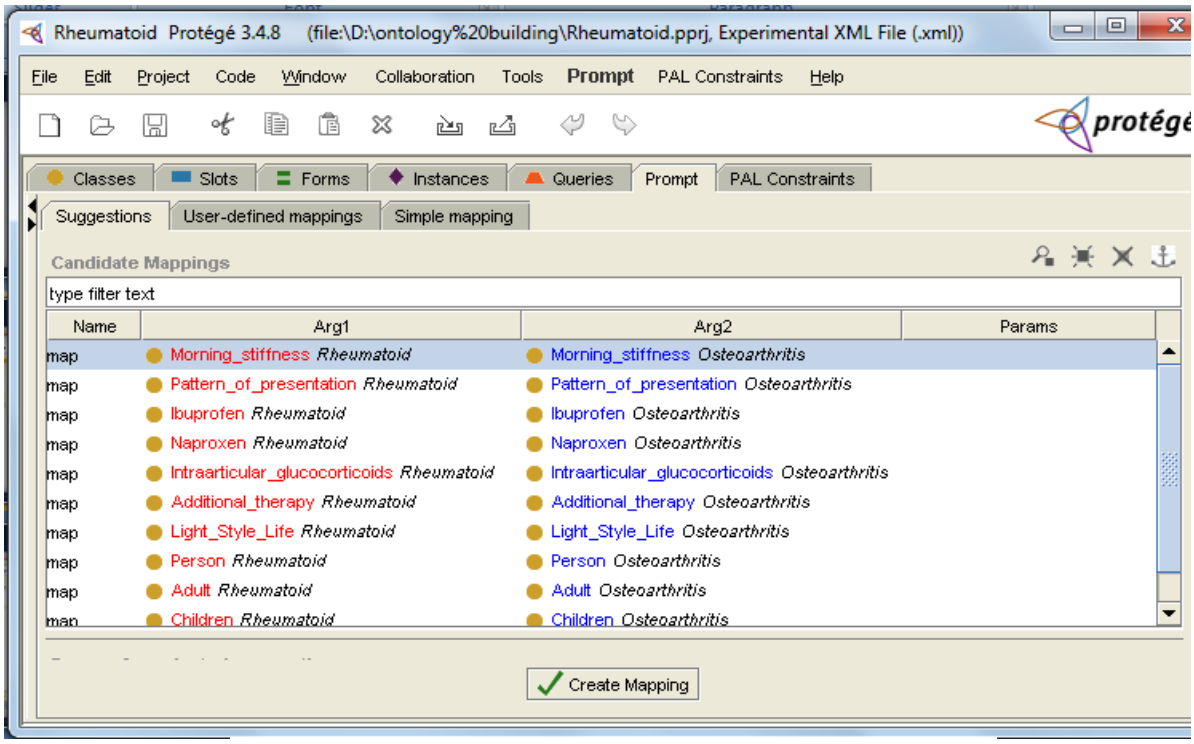

Fig 5 : Results After Using Lexical Matching with Synonym

4. PROPOSAL FOR NEW TOOLS

They are two new tools to propose, the first one is the string equality proposed tool enables the user to import any ontology encode with xml. It uses synonym xml file as an external resource which contains the variants of the two ontologies entities names, the algorithm detects the similar entities regardless of its superclass. While the Smatch can accept the ontology file with xml format only if its classes are written as nodes tags, which require recreating the ontology by the Smatch to apply the matching. The Smatch detects the similarity between entities if they have the same superclass. Although the use of the lexical matching algorithm with synonyms in Prompt Protégé, the results aren't feasible. It can't detect the entities which have the same meaning. The second one is VSMCOS tool which applies the Vector Space Model and cosine similarity as new techniques applied to compute the similarity between two ontology files. Both String Equality Tool and VSMCOS apply the one to many function approach mentioned in section (1).

4.1 String Equality Tool

The tool's technique objective aims to find similar entities (classes and subclasses) between any two ontology files in a specific domain. The similarity in this tool is accomplished to detect any entities in any position which have the same meaning or have the same string structure.

The proposed algorithm begins with the insertion of three files. The first the ontology file1, the second is the ontology file 2, and the last a synoym xml file which contains some terms and its synonym is adapted to the two ontology files domain.

The algorithm used to implement this technique follows these steps:

• It begins by comparing ontology file1 with synonym xml file, the detected similar nodes are stored.

•It Compares ontology file2 with synonym xml file, the detected similar nodes are stored.

•If similar nodes of first step equals to the similar nodes of second step. The algorithm displays them to the user.

•The algorithm checks the string equality between the ontology file1 entities and the ontology file2 entities.

4.2 VSMCOS Tool

The VSMCOS tool applies the Vector Space Model (VSM)[14] and cosine similarity . VSM is widely used in the areas of information extraction and machine learning. It was initially designed as a model to represent arbitrary text documents as vectors from a common vector space [Manning et al., 2008]. It can be adapted to work with any type of data. This technique requires detecting all the entities of each ontology file and storing it in a text document.

The Vector Space Model is executed by the following equations:

- Term Frequency (TF) : is defined as the frequency of a term t appeared in document D .
- The normalized term frequency(NTF): calculates the frequency of a term (entity) appeared in document D divided by the total number of terms in this document.

- Term frequency (TF) -inverse document frequency (IDF): the inverse document frequency of term(t) can be defined using this expression: $[\log(N/(n_j+1))+1]$ where N is the total number of document n_j is the document frequency of term (t). TF-IDF = raw term frequency (t)* IDF (t).
- The cosine similarity (COSSIM): calculates the similarity between two documents.

5. FRAMEWORK PROPOSAL FOR COMPARATIVE STUDY

Ontology matching tools "Smatch and Prompt" are tested by four ontology files; the first two files relates to the biochemistry field, while Rheumatoid and Osteoarthritis relate to the medical field.

Points of comparison that are used to point out differences between Smatch, Prompt and the two proposed tools:

1- Input.

2- Ontology matching technique.

3- Output.

As Table(1) shows the comparison between input. For the Smatch, the ontology files must have xml format and include the xml tags as nodes which require rewriting the ontology file by the Smatch while the other tools can accept the general xml file.

Table1. Tools Comparison With Input

Input	Smatch	Prompt	String Equality tool	VSMCOS Tool
Ontology File Format	Xml	Owl,xml, rdf-xml	Xml	Xml

The Table (2) shows the comparison between used techniques and output for the tools. For Smatch and Prompt, they apply the element level technique 'string-based technique'. The Smatch applies also the language-based method, it used an external lexicon 'Wordnet' to detect the entities that have different structure but have the same meaning. The string equality tool applies the string based technique, the language based method and semantic technique by the use of external terminology file "synonym.xml". The VSMCOS applies the Vector Space Model which could be considered as new technique to measure the similarity between the two ontology files.

For the output, Smatch detects the similarity "equivalence" between entities only if they have the same superclass, and shows other relations "disjoint and more general relations". Although the use of Wordnet it can't detect the entities which have the same meaning. The Prompt output reflects any identical entities that have the same structure. Altough the use of lexical matching with synonym algorithm in Prompt, entities that have the same meaning are not detected.

Table 2. Tools Comparison With Ontology Matching Techniques and Output.

	Smatch	Prompt	String equality tool	VSMCOS tool
Ontology Matching Technique	Applies the string based technique, language based and semantic technique method 'WordNet'	Applies the string based technique.	Applies the string based technique and language based method and semantic technique through the use of 'synonym ..xml'	Applies the Vector Space Model
Output	As figure [Fig.6] shows the similar entities of the two ontology files	As figure [Fig.5] displays the similar entities of the two ontology files	As figure [Fig.7] shows The following : 1- Similar entities of the two ontology files. 2- Entities which have different structure but have the same meaning	As the figure [Fig.8] shows the following : TF NTF. TF.IDF COS-SIM

The figure [Fig.6] shows the results of Smatch Tool, the Rheumatoid and Osteoarthritis ontologies are recreated by the Smatch. The Smatch cannot detect the common class between the two ontologies which is thee "morning _stiffness" class.

The following figure [Fig.7] shows the string equality tool, the imported ontology files are Rheumatoid and Osteoarthritis ontologies. On the left panel, the entities which have the same meaning are displayed an on the right panel, the entities which have the same string structure are displayed.

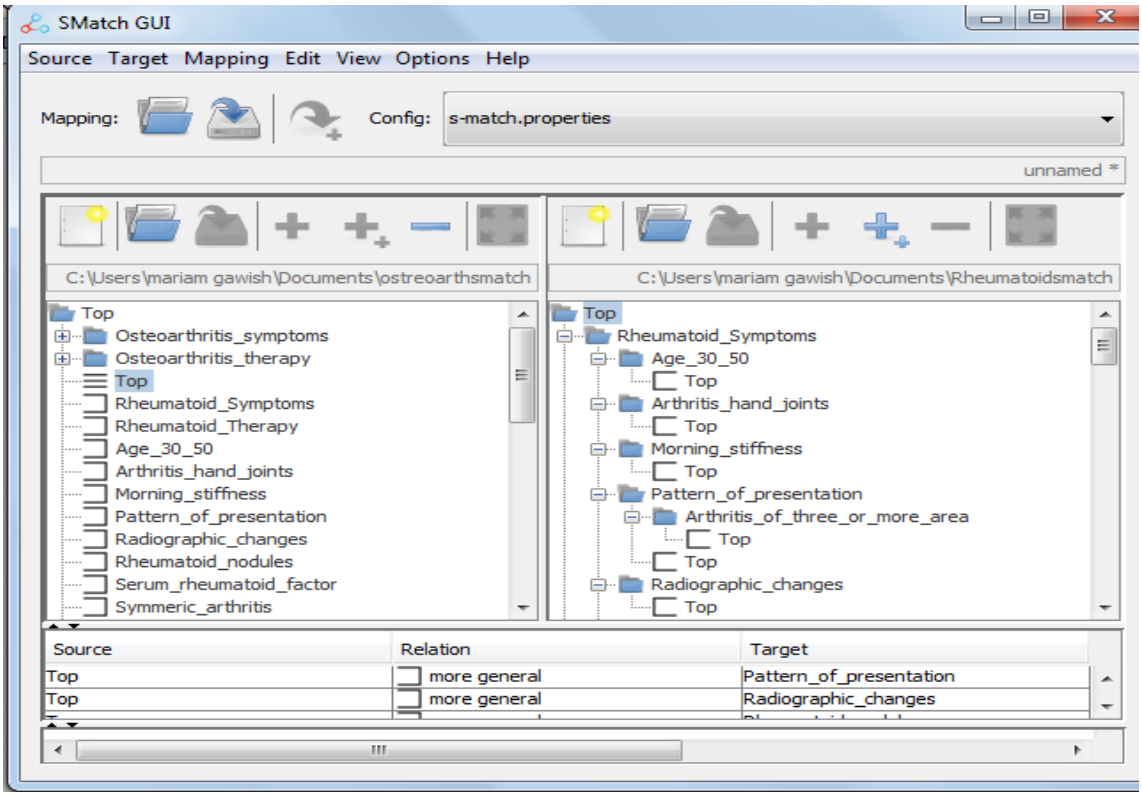

Fig 6: Results of Smatch Tool

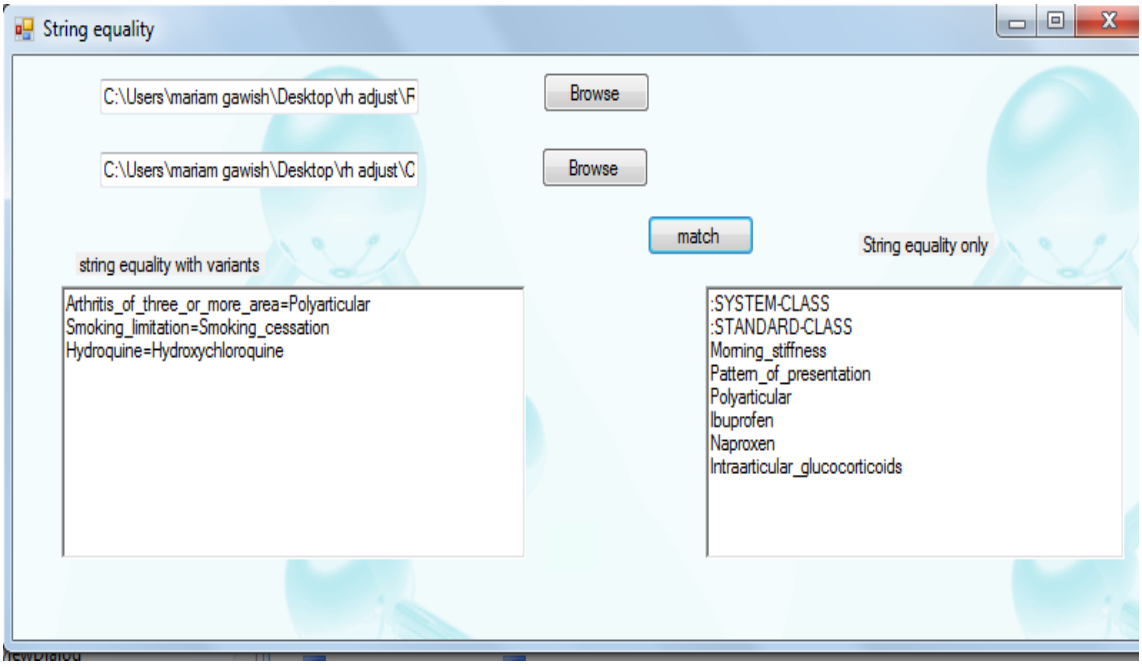

Fig 7: Results of String Equality Tool

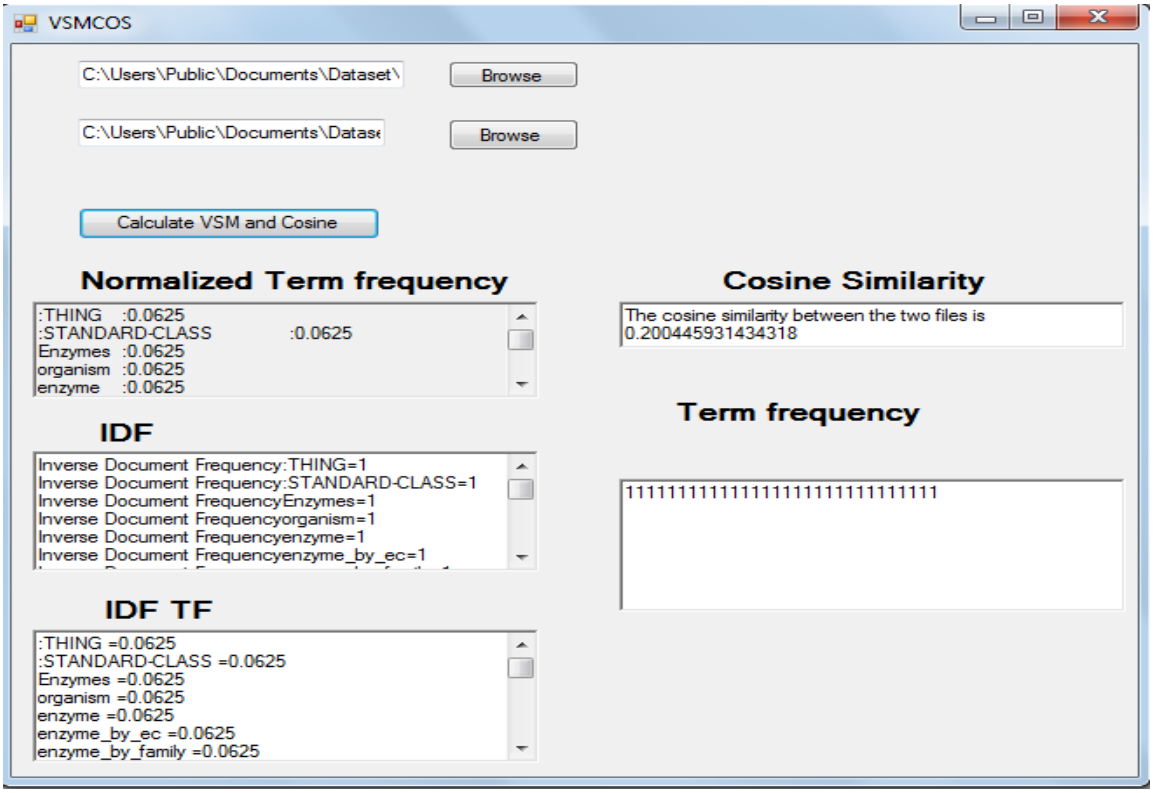

Fig 8 : Results of VSMCOS Tool

The above figure [Fig 8] shows the results of VSMCOS Tool, the imported ontology files are Rheumatoid and Osteoarthritis ontologies. The tool shows the similarity measure between the two ontologies files by calculating the cosine similarity. If the cosine similarity value is high, this indicates that the similarity between the two ontologies files is strong and vice versa.

6. CONCLUSION

Both Smatch and the string equality tool apply the same techniques, but the string equality tool shows better results than the Smatch, which uses the WordNet which doesn't contain the biochemistry and medical terminology. Smatch is suitable to map between the versions of the same ontology e.g.: old version of course ontology and new version of course ontology, but it isn't effective to map between two different ontologies of the same domain.

Prompt and the string equality tools are reliable to match between the different ontologies with different domains.

The VSMCOS applies the Vector Space Model and cosine similarity which shows the similarity ratio between two ontology files. Consequently, it can be considered an indicator which reveals if the two ontologies have the same domain. Hence, the Vector Space Model and cosine similarity techniques can be used as ontology matching preprocess before the application of other ontology matching techniques

7. ACKNOWLEDGEMENTS

I want to thank Dr.Essam Iskandar and Dr.Nabil Chehade for their help to construct the medical ontologies.

8. REFERENCES

[1] Castano, S., Ferrara, A., Montanelli, S., Hess, G.N., Bruno, S.: State of the Art on Ontology Coordination and Matching. BOEMIE Bootstrapping Ontology Evolution with Multimedia Information Extraction Project, FP6-027538 D4.4 (2007).

[2] T. Gruber. Towards principles for the design of ontologies used for knowledge sharing. Int. J. of Human and Computer Studies, 43:907–928, 1994.

[3] J.Euzenat, P.Shvaiko.Ontology matching,pages 9-9,2007.

[4] J.Euzenat, P.Shvaiko.Ontology matching.Page 69-116,2007.

[5] Smatch,http://semanticmatching.org/semanticmatching.html ,retrieved on April 2012.

[6] Wordnet a lexical database for English,http://wordnet.princenton.edu/, retrieved on May 2012.

[7] Y.Liang,H.alani,N.shabolt,Ontology Change Management In Protégé.Intelligence ,Agents and Multimedia Group,School of Electronics and Computer Science,University of Southampton.2005.

[8] GlycO,http://lsdis.cs.uga.edu/projects/glycominds/2006/GlycO.owl, retrieved on December 2011.

[9] OBO: Open Biomedical Ontologies, http://obo.sourceforge.net, retrieved on December 2011

[10] Enzyme ontology,http://lsdis.cs.uga.edu/projects/glycomics/2006/EnzyO.owl,retrieved on December 2011.

[11] I.Mandoiu,R.Sunderraman,A.Zelikovsky.Bioinformatics Research and Applications, 2009.

[12] Arthritis foundation,http://www.arthritis.org/disease-center.php?disease_id =32, retrieved on May 2012.

[13] Medline Plus. A service of the U.S National library of medicine, http://www.nlm.nih.gov/medlineplus/osteoarthritis.html , retrieved on May 2012.

[14] Sannella Salton, Gerard. Introduction to Modern Information Retrieval.McGraw-Hill, 1983.

[15] Protégé 3.4.8, http://protege.stanford.edu/download/protege/3.4/installanywhere/Web_Installers/, retrieved on April 2012.

[16] Ontobuilder2 Automatic schema matching. http://ontobuilder.bitbucket.org/, retrieved on April 2012.

[17] P.Shvaiko , J.Euzenat. Ontology Matching : State of the art and Future Challenges. 2010.