Extracting and Modeling the Semantic Information Content of Web Documents to Support Semantic Document Retrieval

Shahrul Azman Noah[1], Lailatulqadri Zakaria[1] & Arifah Che Alhadi[2]

[1]Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi Selangor MALAYSIA

[2]Department of Computer Science
Universiti Malaysia Terengganu
21030 Kuala Terengganu, Terengganu, MALAYSIA

samn@ftsm.ukm.my, laila@ftsm.ukm.my, arifah_hadi@umt.edu.my

Abstract

Existing HTML mark-up is used only to indicate the structure and lay-out of documents, but not the document semantics. As a result web documents are difficult to be semantically processed, retrieved and explored by computer applications. Existing information extraction system mainly concerns with extracting important keywords or key phrases that represent the content of the documents. The semantic aspects of such keywords have not been explored extensively. In this paper we propose an approach meant to assist in extracting and modeling the semantic information content of web documents using natural language analysis technique and a domain specific ontology. Together with the user's participation, the tool gradually extracts and constructs the semantic document model which is represented as XML. The semantic models representing each document are then being integrated to form a global semantic model. Such a model provides users with a global knowledge model of some domains.

Keywords: ontology, information retrieval, semantic document retrieval, semantic information extraction.

1 Introduction

Accessing and extracting semantic information from web documents is beneficial to both humans and machines. Humans can browse and retrieve documents in a semantically manner whereas machine can easily process such structured representations. Furthermore integrating extracted information from multiple documents can provide users with a global knowledge model of some domains. Due to the structure of human knowledge, the tasks of extracting semantic information in web documents, however, proved to be difficult. The vision of Semantic Web (Berners-Lee et al, 2001) offers the possibility of providing the meanings or semantics of web documents in a machine readable manner. However, the vast majority of 1.5 billion web documents are still in human readable format, and it is expected that this form of representation will still be the choice among content creators and developers due to its simplicity. Due to this phenomenon and the desire to make the Semantic Web vision a reality, two approaches have been proposed (van Harmelen & Fensel, 1999): either furnish information sources with annotations that provide their semantics in a machine accessible manner or write programs that extract such semantics of Web sources.

This research falls into the latter category, whereby the intention is to develop a semi-automated tool meant to assist in extracting and modeling the semantic information content of web documents using the natural language analysis (NLA) technique and a domain specific ontology. In this approach a set of candidate concepts (key phrases or keywords) is automatically extracted from web documents using heuristic rules. Sentences which relate with these concepts are then analyzed and compared with the domain ontology to construct the semantic information content. This process might be performed with the user's participation depending on the domain ontology. Each semantic domain model of a domain is then integrated together to form the global semantic document model. The approach discussed here might be very much similar to another of our work on semantic document retrieval (Noah et al, 2005). However, the focus of this paper is more on the extraction aspect to represent the semantic information of a web document.

This paper is organized into the following sections. The next section provides the background and related research. Section 3 explains the approach employed in extracting and modeling of the semantic information content of web documents. Section 4 and 5 respectively present the testing results and the conclusion that can drawn from our work.

2 Background and Related Research

The aim of information extraction (IE) is to collect the information from large volumes of unrestricted text. IE isolates relevant text fragments, extracts relevant information from the fragments, and pieces together the targeted information in a coherent framework (Cowie & Lehnert, 1996). IE problems have been popularly deal with NLA techniques. However, a few research have considered using domain ontology (Uren et al., 2006; Villa et al., 2003). We provide some background knowledge of NLA and ontology; and then proceed with some related works.

2.1 NLA and the Semantic Web

NLA is the study of understanding human natural language such that it can be understood and correctly processed by machines. Within the vision of Semantic Web, although, NLA contributions are not directly explicated (Berners-Lee et al., 2006), the technology can do play an important in this very slow but progressing semantic technology. NLA for instance can automatically create annotations from unstructured text that provides data which semantic web applications require (Pell, 2007). Research has also been done on providing an NLA type of interface for describing a semantic content which is then translated into one of the Semantic Web enabling technology i.e. Resource Description Framework (RDF) and Web Ontology Language (OWL) (Schwitter, 2005). Another potential application of NLA to Semantic Web is in terms of annotation. Interestingly there are two very different types of annotation process involving NLA, one is annotating natural language document such HTML documents with a pre-specified domain ontology (Vargas-Vera et al., 2002) and the other is annotating document (can be natural language documents or semantic web documents) with natural language (Katz & Lin, 2002). The work by Vergas-Vera et al. (2002) involves pre-processing of HTML documents and semi-automatically annotate the identified concepts with domain ontology. They developed a tool called MnM. The work by Katz and Lin (2002) on the other hand allow users to augment RDF schema with natural language annotations to in order to make RDF more friendly to human instead of machine alone. The work reported in this paper falls into the earlier approach.

2.2 Ontology

Ontologies are widely used in knowledge engineering, artificial intelligence; as well as applications related to knowledge management, information retrieval and the semantic web. Among the first definition of ontology was provided by Neches et al. (1991); which said that "*an ontology defines the basic terms and relations comprising the vocabulary of a topic areas as well as the rules for combining terms and relations to define extensions of the vocabulary*". However, the most quoted definition of ontology is based from Gruber (1993); which defined it as "*an explicit specification of a conceptualization*". Although, there has been no universal consensus for the definition of ontology; the aim of ontology is very clear as put forward by Gomez-Perez et al. (2004) that is "*to capture consensual knowledge in a generic way, and that they may be reused and shared across software applications and by groups of people*".

Ontology is considered as the backbone for the Semantic Web and received great attentions from researchers working in this area. Ontology has also been gradually seen as an alternative to enhance information retrieval task. However, the majority of efforts in information retrieval are limited to query expansion and relevance feedback by exploiting the so-called linguistic ontology such as the WordNet (Miller, 1995). In this paper, we extend the use of ontology into a mediator for mapping concepts extracted from documents and to establish the semantic relationships among the concepts.

2.3 Related Work

We briefly discuss three research works which are very related to ours, which are the work by Embley et al. Embley (2004) and Embley et al. (1999); Brasethvik and Gulla (2001, 2002) and Alani et al. (2003).

The work by Embley (2004) use object relationship model, data frames and lexicon (which forms an ontology) to assist in data extraction from web documents. Brasethvik and Gulla (2001, 2002) employ NLA technique and a conceptual model to support the task of document classification and retrieval. In this case, the conceptual model is constructed by a committee from a set of sample documents by identifying the concepts and relationships. This model is then used for classification and retrieval. Alani et al. (2002) on the other hand develop Artequakt which is an information extraction system for extracting information about artist and artifacts using a domain ontology.

Our work is towards the development of a framework for extracting and modeling the semantic information content of web documents. Our work differs from those of Embly which mainly concerns on extracting 'data' that can be queried similar to SQL-like statement. Our work also differs to the work of Alani et al. (2003) which basically concerns with semantic annotation of web documents. Our work, however, share some similar 'motivation' with the work Brasethvik and Gulla (2001, 2002). The difference is that instead of using a conceptual model, we used existing domain ontology and allow interactive user-tool refinement in constructing the semantic model.

3 The Approach

Our approach to semantic information extraction of web documents involves constructing a semantic document model representing each processed documents. To support this process, the approach employs the natural language analysis (NLA) technique and a set of domain specific ontology. Both are used to perform the task of textual analysis which results not only in the identification of important concepts represented by the documents but also the relationships between these concepts. This approach, therefore, follows the general approach to building semantic index as illustrated by Desmontils and Jacquin (2001).

Figure 1 illustrates the overall process involved in constructing the semantic document model. As compared to the Brasethvik and Gulla (2001, 2002) approach which relies on the conceptual model previously constructed by a selective of personnel, our approach utilise existing ontology of the chosen domain, i.e. the medical domain. A detail discussion of the process therefore follows.

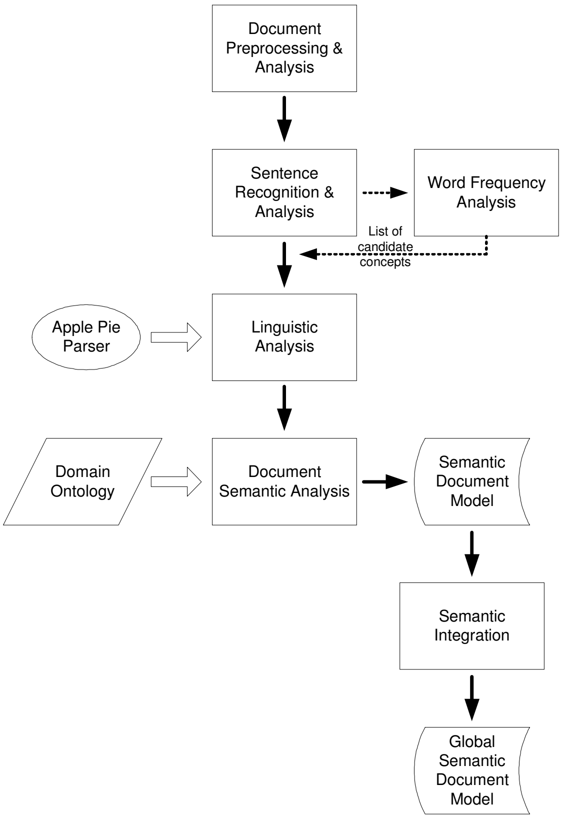

Document Preprocessing & Analysis
Sentence Recognition & Analysis
Word Frequency Analysis
List of candidate concepts
Apple Pie Parser
Linguistic Analysis
Domain Ontology
Document Semantic Analysis
Semantic Document Model
Semantic Integration
Global Semantic Document Model

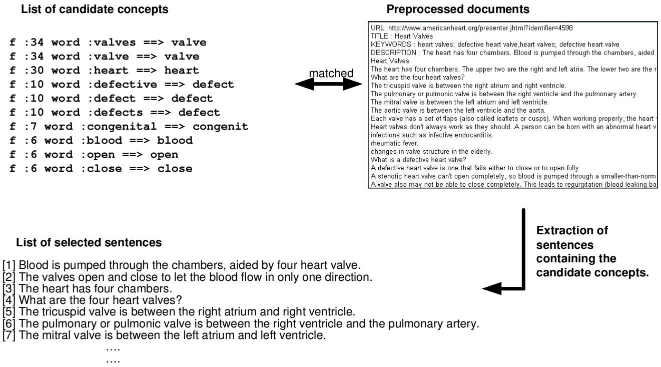
Figure 1. The construction of semantic document model

3.1 Document Pre-processing and Analysis

Every HTML documents sent for processing will firstly be decoded to generate ASCII document files type that are free from any HTML tags. The documents' content extracted from this process are the document's metadata such as the document's title, URL, description and keywords. The textual content of the documents is also extracted.

These documents will then undergo a word analysis process which involved document's filtering and words frequency calculation. In document's filtering, all stop words will be eliminated and selected concepts will be stemmed to their root words. These concepts or words will be sorted according to the frequency of appearance within the document. The sentence analysis and recognition on the other hand will divide the documents into paragraphs, which are in turn broken down into sentences and stored in the document sentences repository. A set of concepts with high frequency previously obtained from the word analysis process will be matched with the sentences stored in the repository in order to select the candidate sentences to be used in the next NLA process. According to Luhn (1958), extracted words with high frequency can represent document's content. The selection of sentences that contains high frequency concepts is entirely based on the heuristic which suggest that such sentences are best describe the content of documents. This heuristic also removes the needs to analyse all possible sentences found in the document which can jeopardize the processing performance of the system.

An example of a document pre-processing and analysis process is as illustrated in Figure 2. As can be seen, the results of this process is a list of potential candidate concepts (for building the semantic document model) sorted according to the number of occurrences – as well as a list of potentially rich information sentences in which the candidate concepts were found.


List of candidate concepts
Preprocessed documents
matched
List of selected sentences
Extraction of sentences containing the candidate concepts.

```
f :34 word :valves ==> valve
f :34 word :valve ==> valve
f :30 word :heart ==> heart
f :10 word :defective ==> defect
f :10 word :defect ==> defect
f :10 word :defects ==> defect
f :7 word :congenital ==> congenit
f :6 word :blood ==> blood
f :6 word :open ==> open
f :6 word :close ==> close
```

```
URL :http://www.americanheart.org/presenter.jhtml?identifier=4598
TITLE : Heart Valves
KEYWORDS : heart valves; defective heart valve,heart valves; defective heart valve
DESCRIPTION : The heart has four chambers. Blood is pumped through the chambers, aided
Heart Valves
The heart has four chambers. The upper two are the right and left atria. The lower two are the r
What are the four heart valves?
The tricuspid valve is between the right atrium and right ventricle.
The pulmonary or pulmonic valve is between the right ventricle and the pulmonary artery.
The mitral valve is between the left atrium and left ventricle.
The aortic valve is between the left ventricle and the aorta.
Each valve has a set of flaps (also called leaflets or cusps). When working properly, the heart
Heart valves don't always work as they should. A person can be born with an abnormal heart v
infections such as infective endocarditis.
rheumatic fever.
changes in valve structure in the elderly.
What is a defective heart valve?
A defective heart valve is one that fails either to close or to open fully.
A stenotic heart valve can't open completely, so blood is pumped through a smaller-than-norm
A valve also may not be able to close completely. This leads to regurgitation (blood leaking ba
```

[1] Blood is pumped through the chambers, aided by four heart valve.
[2] The valves open and close to let the blood flow in only one direction.
[3] The heart has four chambers.
[4] What are the four heart valves?
[5] The tricuspid valve is between the right atrium and right ventricle.
[6] The pulmonary or pulmonic valve is between the right ventricle and the pulmonary artery.
[7] The mitral valve is between the left atrium and left ventricle.
....
....

Figure 2. An example of the document analysis process

3.2 Natural Language Analysis

The natural language analysis process can be divided into two subsequent stages: the morphology and syntactic analysis; and the semantic analysis. The main aim of this process is to generate a local semantic document model representing each processed document The morphology and syntactic analysis process will analyse the input sentences previously stored (sentences that contains candidate concepts) in the sentences repository into a parse tree using the Apple Pie Parser (Sekine, 2006). The parser is a bottom-up probabilistic chart parser which finds the parse tree with the best score by best-first search algorithm. The grammar used is a semi context sensitive grammar with two non-terminals and was automatically extracted from Penn Tree Bank, syntactically tagged corpus made at the University of Pennsylvania (Sekine, 2006).

The process of morphology and syntactic analysis is considered to be domain independent. For example the input sentence of "*Blood is pumped through the chambers, aided by four heart valves*", is being parse to the following parse tree.

(S
  (NPL *Blood*)
  (VP *is*
    (VP *pumped*
    (PP *through*
      (NP
        (NPL *the chambers*) -COMMA-
        (VP *aided*
          (PP *by*
            (NP
              (NPL *four heart*) *valves*)))))))
                              -PERIOD-)

The semantic analysis on the other hand performs the task of extracting the semantic relationships between the selected concepts. This is perform either by the use of domain specific ontology or by exploiting the semantic structure of the analysed sentences with the help of the user. The interactions from user at this stage is seen acceptable as fully-automated approach to semantic analysis is not possible due to the requirement for deep understanding of the domain in concern (Snoussi et al., 2002). Figure 3 illustrates the overall process of this stage which indicates that the two main activities involved are the identification of concepts and the relationships between these concepts. Detail discussion of these activities therefore follows.

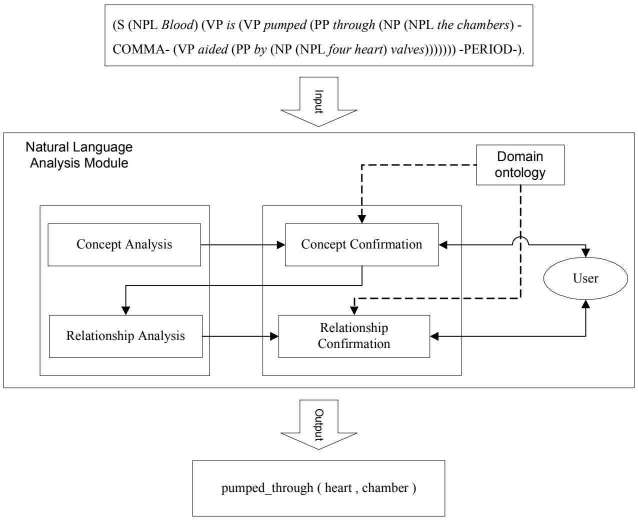

(S (NPL Blood) (VP is (VP pumped (PP through (NP (NPL the chambers) -COMMA- (VP aided (PP by (NP (NPL four heart) valves))))))) -PERIOD-).
Input
Natural Language Analysis Module
Domain ontology
Concept Analysis
Concept Confirmation
User
Relationship Analysis
Relationship Confirmation
Output
pumped_through (heart , chamber)

Figure 3. The natural language analysis process

Noun phrases and verb phrases are good indications of concepts to be included in the semantic documents model. Therefore, every noun phrases and verb phrases extracted from the analysed sentences are represented as concepts. These noun phrases will be analysed to filter determiners (such as *the*, *a* and *and*) that usually occur in word phrases.

For example, the parsed sentences of *"Blood is pumped through chambers aided by four heart valves*" in the form of (S (NPL *Blood*) (VP *is* (VP *pumped* (PP *through* (NP (NPL *the chambers*) -COMMA- (VP *aided* (PP *by* (NP (NPL *four heart*) *valves*))))))) -PERIOD-), will resulted in the extraction of the concepts: '*blood*', '*the chambers*' and '*four heart'*. The determiner '*the*' and the stopword '*four*' will be removed from the identified concepts.

The confirmation (in terms of the correctness) of the filtered concepts is performed in two ways; either automatically endorsed by referring to the domain ontology if the mapping between the concepts and the domain ontology existed; or from user intervention in the case where no such mapping found existed. Figure 4 illustrates a portion of the domain ontology used by the tool.

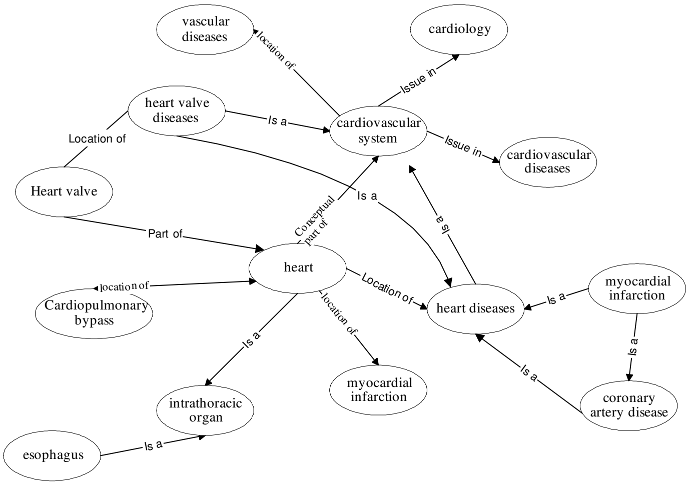


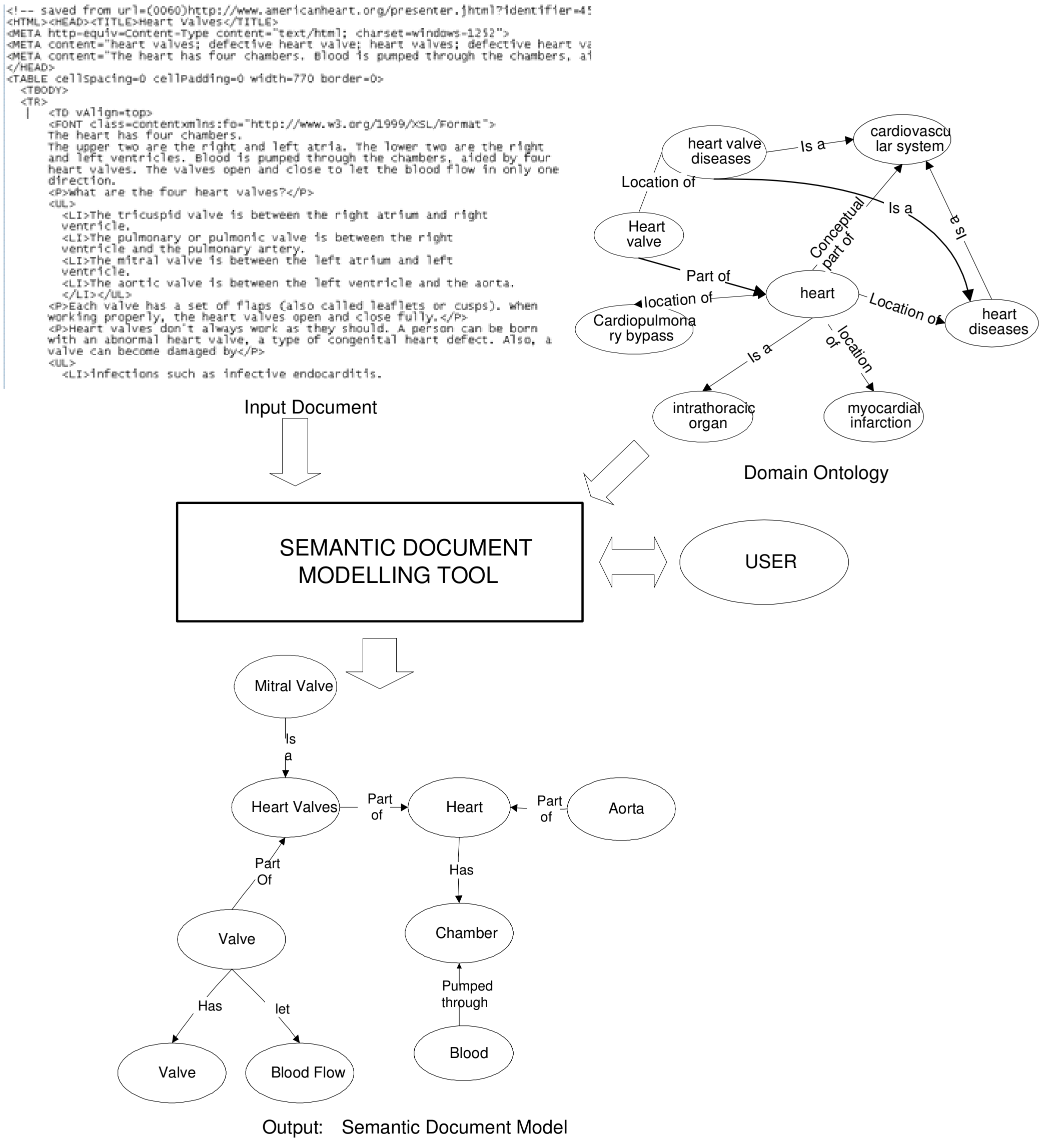
Figure 4. A segment of the heart domain ontology

Relationship recognition identified during the previous phase (concept recognition and confirmation) against concepts within the is done by comparing candidate concepts and concepts which were domain ontology. If a pair of concepts found matched with the domain ontology, the relation of these concepts is automatically defined by referring to the domain ontology if such a relation existed. If a relation does not exist, a suggestion is provided based upon the syntactic sentence structure of the associated concepts, of which the user will define it manually. Similarly, for those concepts not presented in the domain ontology, the tool will first provide a list of concept candidates which can best be linked based upon the analysis of the chosen sentences. Once the desired concepts have been selected, the tool will provide the suggestion of possible relationships between these concepts.

Figure 5 illustrates an example of a HTML document, a fraction of medical domain ontology and the output generated by the semantic document modeling tool. As can be seen from this example, the semantic relationships of "*mitral valve part-of heart*", "*heart valve part-of heart*" and "*mitral valve is-a heart*" are all extracted from the domain ontology whereas the other concepts and relationships are extracted by means of text analysis with the user's participation. The generated semantic document model is an XML representation of the concepts, relationships as well as the URL of the selected documents. Example below is part of the generated XML representation. This model is then stored in the Semantic Document Model.

```
<?xml version="1.0" encoding="UTF-8" ?>
<DocumentInfo>
<MetadataInfo>
<Title>Heart Valves</Title>
<Url>http://www.americanheart.org/presenter
.jhtml?</Url>
<Keywords>heart valves , heart , mitral
valve , aorta , blood , chambers , valves ,
blood flow , valve , flaps ,</Keywords>
</MetadataInfo>
<Semantic_Content>
<Concept>
   <ConceptDescription>
```

```
      <String>heart valves</String>
   </ConceptDescription>
   <ConceptRelationship>
      <part_of>
         <String>heart</String>
      </part_of>
   </ConceptRelationship>
</Concept>
<Concept>
   <ConceptDescription>
      <String>mitral valve</String>
   </ConceptDescription>
   <ConceptRelationship>
      <is_a>
         <String>heart valves</String>
      </is_a>
      <part_of>
         <String>heart</String>
      </part_of>
   </ConceptRelationship>
</Concept>
<Concept>
   <ConceptDescription>
      <String>aorta</String>
   </ConceptDescription>
   <ConceptRelationship>
      <part_of>
         <String>heart</String>
      </part_of>
      </part_of>
   </ConceptRelationship>
</Concept>
………………
……………. . .
</Semantic_Content>
```



```
<!-- saved from url=(0060)http://www.americanheart.org/presenter.jhtml?identifier=45
<HTML><HEAD><TITLE>Heart Valves</TITLE>
<META http-equiv=Content-Type content="text/html; charset=windows-1252">
<META content="heart valves; defective heart valve; heart valves; defective heart va
<META content="The heart has four chambers. Blood is pumped through the chambers, ai
</HEAD>
<TABLE cellSpacing=0 cellPadding=0 width=770 border=0>
  <TBODY>
  <TR>
  |   <TD vAlign=top>
      <FONT class=contentxmlns:fo="http://www.w3.org/1999/XSL/Format">
      The heart has four chambers.
      The upper two are the right and left atria. The lower two are the right
      and left ventricles. Blood is pumped through the chambers, aided by four
      heart valves. The valves open and close to let the blood flow in only one
      direction.
      <P>What are the four heart valves?</P>
      <UL>
        <LI>The tricuspid valve is between the right atrium and right
        ventricle.
        <LI>The pulmonary or pulmonic valve is between the right
        ventricle and the pulmonary artery.
        <LI>The mitral valve is between the left atrium and left
        ventricle.
        <LI>The aortic valve is between the left ventricle and the aorta.
        </LI></UL>
      <P>Each valve has a set of flaps (also called leaflets or cusps). When
      working properly, the heart valves open and close fully.</P>
      <P>Heart valves don't always work as they should. A person can be born
      with an abnormal heart valve, a type of congenital heart defect. Also, a
      valve can become damaged by</P>
      <UL>
        <LI>infections such as infective endocarditis.
```

Figure 5. Constructing the semantic document model

3.3 Integration of Semantic Document Model

The stored semantic document models will then undergo an integration process which results in the creation of a global semantic document model. The global semantic model is meant to be used for semantic retrieval and browsing.

The semantic integration is an uncomplicated process requiring insertion of a new semantic document model to the existing global semantic document model. The process will remove aspects of redundancies and documents that belong to the same semantic concept are clustered together. A set of simple object type and mismatch rules which was mainly derived from the theory and technique of automatic conceptual modeling integration process (Batini et al., 1986); Noah & Williams, 2004) have been used. The rules are mainly for binary relationships as the semantic document model represented are binary in nature. At the moment aspects pertaining to the concepts of generalization, aggregations and associations of a semantic model are being considered. Naming conflicts such as synonyms and homonyms however are not being considered by the rules.

4 Results

Evaluation was done by comparing the extracted concepts in semantic document model with keywords in <META> tag provided by authors. Our assumption was that <META> tags provide key information or phrase reflecting document content. A similar method of evaluation has been conducted by Witten et al. (2000) and Song et al. (2004) which compare the extracted concepts with those human-generated keywords or key phrase in order to evaluate the performance of KEA and KPSpotter respectively.

The main inherent problem with this evaluation approach is the lack of web pages that provide <META> tags keywords which resulted in the limited number of available testing document collection. We have performed hundreds of document analysis but only 50 web documents were acceptable and sufficient enough for testing. Table 1 shows the result of average 'correct concepts' corresponding to the concepts assign by author (extracted from <META> tags).

Table 1 lists the average number of matched candidate concepts, system extracted concepts and concepts from the generated semantic document model for the 50 test web documents. Candidate concepts referred to concepts used for selecting potentially rich information sentences during the document analysis process. Ontological-based extracted concepts are concepts extracted from the domain ontology and used to generate portion of the semantic document model. The semantic document model concepts are concepts presented in the final generated semantic document model. In other words the semantic model concepts are the composition of matched concepts derived from the domain ontology and concepts confirmed from the activities of user interactions.

Concepts extracted	Average number of concepts matched with <META> tags		
	Candidate Concepts	Ontology-based extracted concepts	Concepts in generated semantic document model
1	0.75	0.52	0.56
2	1.6	1.14	1.18
3	2.32	1.48	1.64
4	3.14	1.66	1.96
5	4.18	1.84	2.26
6	4.93	1.92	2.48
7	5.93	1.98	2.76
8	6.72	2.06	3.12

Table 1: Overall Performance.

As can be seen, the average numbers of system extracted concepts with user interactions (that corresponds to the concepts assigned by authors) is 3.12. Therefore, system extracted concepts with user interventions capable of extracting one to three 'correct concepts'. System-extracted concepts achieved one to two correct concepts which is 2.06 in average. System extracted concepts depends fully on domain ontology for concepts identification and extraction limits to the stored information. Our implementation of domain ontology only stores 24 related concepts which actually represent a small fraction of the 22,997 terms listed in the Medical Subject Heading (MeSH) Concepts assigned by authors cover a wider range of concepts in domain ontology and sometimes go beyond the domain ontology itself. Adding more information/concepts into domain ontology may increase system efficiency in performing concepts identifications. System extracted concepts with user intervention achieved a better result because it allows user to describe web document content based on their understanding of the domain.

Based on the testing, our approach capable of extracting at between one to six correct candidate concepts. In other words for the first eight concepts an average of six concepts matched with those of authors' concepts (i.e. meta tags). However, for every eight concepts extracted of which six were selected as candidate concepts, only about three concepts were presented in the generated semantic document model. This difference is resulted by the following possibilities.

- The candidate concepts do not matched with any of the ontological domain concepts.
- The candidate concepts and the ontological domain concepts were not used to construct the semantic document model because their presence are not inherent in the filtered sentences.

Having one to three 'correct concepts' in the generated model does not indicate that other concepts are not representative of the domain. Our testing result, however, is higher than those reported by KEA and KPSpotter that respectively shows 1.8 and 2.6 extracted key phrases in average that match with author key phrases. This technique of evaluation, however, does not consider semantic relationships between extracted concepts. Such 'correctness' of relationship is best judge independently by human beings.

The results of our testing might also be influenced by the way authors describe their respective documents with <META> tags, which may be summarize as follows:

- Authors do not always choose the best keywords or key phrases to reflect the content of their documents. In some cases, authors apply the same set of keywords for different documents.
- Authors, instead of using the same keyword or key phrases that they use in their documents, they replace them with other concepts which are similar or synonyms. As a result, some of those keywords are not available within their document.

4 Conclusions and Future Works

Domain ontology plays an important role in supporting the tasks of document classification and organization. In this paper, we have presented how a domain ontology combine with a natural language analysis technique can be exploited not only to extract important concepts from documents but also to construct the semantic content of web documents.

Although, controlled vocabulary has been used in information retrieval systems (Embley, 1999), the vocabulary tends to be a list of terms that are syntactically matched with terms in documents. The inherent meanings or structures of the terms in the vocabulary are not used to represent the semantic meanings of documents, and users are still left with a syntactic approach to information retrieval. While ontologies for Semantic Web have been focus to support machines looking for information instead of human, semantic document model is intended to support human communication, which requires a human readable notation. In our case the constructed semantic document model is rather meant for later retrieval by human instead of machines or software agents. Our current research work is to enhance aspects of global semantic document integration by considering further integration aspects such as synonyms, homonyms and inheritance mechanisms which are very well established within the context of database conceptual modeling. Testing and evaluation of the approach presented in this study are also currently being carried out.

4 References

T. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web. *Scientific American*, May, pp. 35-43, 2001.

F. van Harmelen, and D. Fensel, Practical knowledge representation for the Web. *IJCAI Workshop on Intelligent Information Integration*, 1999.

S. A. Noah, A. C. Alhadi. and L. Zakaria, A semantic retrieval of web documents using domain ontology. *International Journal of Web Grid and Services,* pp. 151–164, 2005.

E. Desmontils and C. Jacquin, Indexing a web site with terminology oriented ontology. *International Semantic Web Working Symposiums (SWWS)*, Stanford University, California, 2001.

J. Cowie and W. Lehnert, Information Extraction. *Communications of the ACM*, 39(1), pp. 80-91, 1996.

V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta and F. Ciravegna, Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4, pp. 14-28, 2006.

R. Villa, R. Wilson, and F. Crestani, Ontology mapping by concept similarity. *International Conference on Digital Libraries*, pp. 666-674, 2003.

T. Berners-Lee, W. Hall, J.A. Hendler, K. O'Hara, N. Shadbolt, N. and D.J. Weitzner, A Framework of Web Science. *Foundations and Trends in Web Science*, 1(1), pp 1-25, 2006.

B. Pell, POWERSET - Natural Language and the Semantic Web. *The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, 2007.*

R. Schwitter, A Controlled Natural Language Layer for the Semantic Web. *AI 2005: Advances in Artificial Intelligence*, pp. 425-434, 2005.

M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna, MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. *Proc. of EKAW 2002*, pp. 379-391, 2002.

B. Katz, and J. Lin, Annotating the Semantic Web Using Natural Language. *Proceedings of the 2nd Workshop on NLP and XML, Taipei, September* 200, pp. 1-8, 2002.

R. Neches, R.E. Fikes, T. Finin, T.R. Gruber, T. Senator and W>R. Swartout, Enabling Technology for Knowledge Sharing. *AI Magazine* 12(3), pp. 36-56, 1991.

T.R. Gruber, A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition* 5(2), pp. 199-220, 1993.

A. Gomez-Perez, M. Fernandez-Lopez and O. Corcho, Ontological Engineering. Berlin: Springer-Verlag, 2004.

G. Miller, WordNet: A Lexical Database for English. *Communications of the ACM* **38**(11): pp. 39-41, 1995.

D.W. Embley, Toward Semantic Understanding – An Approach Based On Information Extraction Ontologie. *Proceedings of the 15th Australasian Database Conference, 2004, (ADC'04)*, Dunedin, New Zealand, pp. 3-12, 2004.

D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.K. Ng and R.D. Smith, Conceptual-model-based data extraction from multiple-record Web pages. *Data and Knowledge Engineering*, pp. 227-251, 1999.

T. Brasethvik, and J.A. Gulla, A Conceptual Modelling Approach to Semantic Document Retrieval. *Advanced Information Systems Engineering, 14th International Conference*, pp.167-182, 2002.

T. Brasethvik and J.A. Gulla Natural language analysis for semantic document modeling. *Data and Knowledge Engineering*, pp. 45-62, 2001.

H. Alani, S. Kim, D. Millard, M. Weal, W. Hall., P. Lewis, P. and N. Shadbolt, Ontology knowledge extraction from web documents. *IEEE Intelligent Systems*, 18(1), pp. 14-21, 2003.

H.P. Luhn, The automatic creations of literature abstracts. I.B.M. Journal of Research and Development, 2(2), pp. 159-165, 1958.

S. Sekine *Proteus Project - Apple Pie Parser (Corpus based Parser).* http://nlp.cs.nyu.edu/app (accessed on 15 September 2006)

S. Sekine and R.A. Grishman, Corpus-based Probabilistic Grammar with Only Two Non-terminals, *Fourth International Workshop on Parsing Technology*, pp. 216-223, 1995.

H. Snoussi, L. Magnin and J.Y. Nie, Towards an ontology-based web data extraction. *The AI-2002 Workshop on Business Agents and the Semantic Web (BASeWEB),* pp. 26-33, 2002.

C. Batini, M. Lenzerini and S.B. Navathe, A comparative analysis of methodologies for database schema integration, *ACM Computing Surveys,* 18(4), pp. 323-364, 1986.

S.A. Noah and M. Williams, Intelligent object analyzer for conceptual database design, *Jurnal Teknologi*, 3, pp. 27-44, 2004.

I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin and C.G. Nevill-Manning, KEA: Practical automatic keyphrase extraction, *Working Paper 00/5*, Department of Computer Science, The University of Waikato, 2000

M. Song, I.Y. Song and X. Hu, An efficient keyphrase extraction system using data mining and natural language processing techniques. *First International Workshop on Semantic Web Mining and Reasoning*, pp. 58-65, 2004.