Chapter 2

Literature Survey

The objectives of the research have been identified after performing an exhaustive review of the literature. This chapter is divided into a few sections in order to clearly focus on the contributions of various researchers under each area. The first section discusses about the research contributions carried out in creating a knowledge base using ontology. The second section discusses the work of researchers and the different methods followed in automating an ontology from the resources. The third section discusses about the mining algorithms and their use in concept extraction from the resources. The fourth section discusses the need for semantic technologies and ontology in e-learning systems. The last section deals with the classification of similarity measures of single ontology and multiple ontologies. A brief discussion about the literature of single ontology similarity measures and the classifications of hybrid measures are discussed in detail.

2.1 KNOWLEDGE EXTRACTION USING ONTOLOGY

Lots of information exists in various databases and World Wide Web (WWW) repositories, but extracting knowledge from such repositories was a tedious task for users. The main problem was to organize or group the relevant information to acquire the required knowledge relating to a particular topic. Knowledge is represented in the form of structured and unstructured sources. A few structured sources are eXtensible Markup Language (XML), relational databases etc. and unstructured sources are texts, documents, presentations and images. When there are varied sources of information in the repositories, knowledge extraction and information retrieval requires a large number of searches. One of the best approaches to simplify the search in Web is to organize the existing knowledge in a standard machine understandable format. This can be achieved with the help of semantic web that provides information access based on machine processable data (Dieter Fensel and Musen, 2001). The explicit
representation of data in a standard format is done using semantic web languages like RDF, OWL which provides an XML like syntax (Michel, 2001). Harith Alani et al. (2003) have designed an Artequakt project which generated the biographies of artists which was spread across the Web. Techniques were discussed for automatically extracting knowledge about artists from Web documents by applying information extraction tools from online documents based on a given ontology representation. The ontology for this work was constructed using the sections of a standard ontology called Conceptual Reference Model Ontology which was developed to represent cultural heritage information. This ontology was modified for the Artequakt project to store the artist details, their personal information, artwork details, and family details etc. All this information was stored in knowledge base and analysed for redundant facts. WordNet was used to reduce linguistic variations that exist between the relations in the ontology and in the text extracted. To arrange the details of the artist as a biography, a template or structure was designed by a human as per the requirements and stored as an XML file. This template was used to generate the biographies of the artists.

The involvement of a semantic web can enhance the ability of users in document retrieval and finding more relevant answers to the queries on Web. Tim Finin et al. (2005) have discussed the use of semantic web in information retrieval. Already there are lots of researches (Arocena and Mendelzon, 1998 ; Bar-Yossef et al. 1999 ; Egnor and Lord, 2000 ) going on based on structuring data and information retrieval. Here a framework was proposed which would support both retrieval and inference based processing. It means indexing can be done using text and markup terms which was not possible in conventional search engines. Swangling technique was used to encode the markup query to text query. This was tested using three prototype systems called OWLIR, Swangler and Swoogle. The need for integrated knowledge technologies and ontologies to facilitate meaningful knowledge acquisition and sharing of information was discussed by Anagnostakis, A.G. et al. (2005). This platform was entitled CITATION and it was based on the healthcare administration sector where there was overwhelming amount of information that needs to be structured and indexed for effective retrieval. The proposed platform automatically structures documents in XML format and indexes them with ontologies which provide information over specific domains. It also supported multilingual
information retrieval. The performance of the system was evaluated based on the recall and precision values obtained.

Another efficient approach for knowledge extraction from textual Web contents was proposed by Tao Jiang et al. (2007). The documents are pre-processed and the terms are extracted based on the sentence grammar trees. The grammar tree is constructed based on a set of predefined rules from a lexical dictionary. The terms are clustered based on the synonyms. Each term extracted from the documents is first checked whether it can join any synonym group, if so, it is added. Else a new group cluster is formed. Each synonym group is encoded as an RDF vocabulary. From the RDF, generalized association rules are mined using GP-close algorithm (Close Generalized Pattern Mining). This algorithm mines rule in a closed pattern instead of all frequent patterns. This experiment was conducted on an online database of terrorist activities. Knowledge discovery is a process that is dependent on the type of application to be processed. Ontologies are created based on the type of applications. By grouping the information based on space and time, effective services can be provided to the users by identifying the user’s location, moving objects and its behaviour, its context and their interest (Mousavi and Hunter, 2012). This can be possible by integrating different ontologies designed for geographic, geometric, theme and service. Proactive services can be provided to the users by extracting patterns based on the context and the behaviour. Recent researches concentrate on using ontologies for knowledge extraction from medical domain and images (Priyamvada et al. 2013). Low level feature extraction methods in image processing were combined with the feature based custom designed ontology to retrieve meaningful images from the Web.

This section pertains to a discussion of the researches on knowledge extraction from documents, done based on a predefined ontology which was developed with the help of human experts. When the size of database or the resources increases, designing an ontology also becomes a tedious process. So the next focus by researchers was on techniques to automate an ontology from the data resources available.
2.2 AUTOMATING ONTOLOGY FROM RESOURCES

This section discusses the ongoing research in the field of ontology automation. There were many semi automated approaches to populate ontology. A method for semi-automatic ontology acquisition from a Corporate Intranet of insurance company was proposed by Keitz J.U. et al. (2000) where GermaNet similar to WordNet was considered as the base ontology. The synsets were converted into concepts if it had a hypernym and hyponym. The relations between the concepts were identified using association rule mining (Apriori) algorithm. This was a partial ongoing work and in its early stage of completion. There have been no papers later published regarding the completion of this work or an extension of this work. Another approach to design a semi automatic framework to derive concept maps and from these concept maps a domain ontology was derived by Amal Zouaq and Roger Nkambou (2009). The main focus of their research was to target the e-learning community and tailor the needs of learners. A tool called TEXt–Concept–Map–Ontology (TEXCOMON) was designed to achieve this goal. A methodology was formulated to convert text documents to concept maps and from there to how it is transformed into an OWL ontology. Two handbooks from Sharable Content Object Reference Model (SCORM) standard were used to create a dataset for the research. The ontologies derived using TEXCOMON were compared with the ontology using OnTo text tool. Syntactic and structural evaluation of ontologies was done to prove that the ontology derived through TEXCOMON was more efficient than the others. Each phase in the process required a human validation which was considered as a disadvantage to the system. Structural measures were density measure, betweeness measure, class match measure and semantic similarity measure. A total of these four measures were used to rank and sort the ontologies.

A light weight ontology was built to capture and reason the communication inside an organization by Grobelnik M. et al. (2009). This helped to identify the pattern of connectedness of employees within an organization. The data set considered was a set of log files of 800-900 people over a period of nearly 20 months. This approach had two phases – one was data pre-processing phase and the second was ontology modelling and visualization phase. In the data pre-processing phase, E-learning and data transformation were done. In the next phase, the sender’s and
receiver’s email communication was created as a graph, by the application of heuristics and analytical rules. The graph is transformed into a sparse matrix where in the element $i^{th}$ row and $j^{th}$ column is non zero if $i^{th}$ and $j^{th}$ vertices are connected directly or indirectly and zero if it is no way connected. This structure was visualized as a geographical terrain to identify the density of communications.

Another technique for creation of domain ontology from the queries posted in the online discussion fora was proposed by Raymond et al. (2009). The students post their messages in the forum and it is really difficult for the course instructor to go through all the messages and analyse the depth of student’s knowledge. Concepts maps were generated based on the messages posted to online discussion fora. Mutual Information is a method to calculate the dependency between two entities. Based on that, a measure called Balanced Mutual Information (BMI) method was used for concept extraction. An ontology was automated from the concepts extracted. By searching the concept maps, instructors can easily identify the progress of their students and can adjust the pedagogical sequence on the fly. Their experimental results exposed the accuracy and the quality of the automatically generated concept maps as pretty good. The future directions discussed in paper was the use of mining techniques for concept extraction. Another framework for extracting ontology from relational databases was proposed by Peng Liu (2010). A set of rules were framed to convert a relational database to an OWL file and the ontology was automated. The increase in the use of video applications and images in recent years also requires the retrieval of contents which are related and meaningful. So a semantic content extraction that helps the users to query and retrieve the videos was designed by Yilidrim et al. (2013) based on a fuzzy domain video ontology.

Extracting ontologies was done from different sources like textual documents, images, videos, fora etc. All the above research contributions concentrated on automating domain ontologies using analytical rules, heuristics and with human intervention in intermediate steps. The previous research contributions proved that mining algorithms were not used in concept extraction. The proposed future enhancements by Raymond et al. (2009) paved way to use association rule mining algorithms to extract the concepts from the textual documents. Based on the data to be
mined, two mining algorithms were selected based on a few criteria from the existing algorithms. This is discussed in detail in section 2.3.

2.3 USE OF MINING ALGORITHMS FOR CONCEPT EXTRACTION

Association rule mining techniques were considered as a standard technique to discover the association between the items. An efficient algorithm to generate the association rules for the items in a database was proposed by Rakesh et al. (1993). This analysis was done using a supermarket database. This algorithm helps us to identify all the qualitative rules based on a minimum transactional support of 1% and confidence of 50%. (In general, rules discovered where of both qualitative and quantitative ones). To find all the item sets in a single pass was difficult, so an estimation procedure was done to strike a balance between the number of passes and number of item sets. In order to avoid unnecessary passes, pruning techniques were used and buffer management techniques were used to hold the data that may not fit in the memory in a single pass. This helps us avoid loss of data. Later, Rakesh and Srikanth (1994) have proposed two new algorithms Apriori and AprioriTid for the discovery of all association rules between items in large database transactions. The best feature of these two algorithms has been combined into a hybrid algorithm called apriori hybrid that scales with the number of transactions.

Support and Confidence were the best known constraints based on which the association rules were selected. It is generally required to satisfy the minimum threshold of support and confidence as specified by the user. An association rule is defined as $A \rightarrow B$ where $A, B \subseteq I$, where $I$ is set of items. Support is defined as the count of $A \cup B$ divided by the total number of transactions in the database.

$$\text{Support} (AB) = \frac{\text{Support count} (A \cup B)}{\text{Total number of transactions in database}}$$

where support count is defined as the total of number of transactions with the frequent itemset. Confidence is defined as given below

$$\text{Confidence} (A|B) = \frac{\text{Support} (AB)}{\text{Support} (A)}$$
All the algorithms discussed above use the generate and test approaches. First, the candidate item set was generated and later the item set was tested to identify if it was a frequent itemset or not. When the size of data set increased, these algorithms were computationally expensive in terms of time and space.

The mining performance could be improved significantly if the candidate set generation was eliminated. This was proposed by Han et al. (2000) and a new association rule mining technique called FP–growth emerged. First, a frequent pattern tree is constructed for storing all the frequent transactions. Only the frequent items will have nodes in the tree i.e., the more frequently occurring nodes have better chances to share nodes than less frequently occurring nodes. The main difference between the two approaches is that the Apriori algorithm uses bottom up approach for generation of frequent itemsets and FP–growth algorithm uses the divide and conquer and partition based approaches. There are many extensions and improvements to Apriori and FP-growth algorithms. In both these algorithms, the transactions are represented in horizontal data format. Another association rule mining algorithm called the Equivalent Class Transformation (ELCAT) algorithm (Zaki, 2000) was used to mine frequent patterns. Here the transactions are represented in vertical format and the computation is done using depth first search. Many comparative studies show that FP-growth algorithm requires less number of scans, memory utilization and time to find the frequent patterns than all other association rule mining algorithms.

Another mining technique where the transactions were ordered elements or a subsequence of patterns was introduced by Agarwal and Srikanth (1995). Every transaction can be recorded with or without the timestamp depending upon the dataset. Generalized sequential pattern mining (GSP) (Srikanth and Agarwal, 1996) was an extended version of apriori based sequential pattern mining. Zaki, (2001) developed another sequential pattern mining algorithm called SPADE (Sequential pattern discovery using equivalence classes) which was an extension of ELCAT algorithm. The sequence of patterns is divided based on sub sequences and the sequence database is projected based on the partition of patterns. Such a pattern mining is called sequential pattern mining by pattern growth. PrefixSpan (Pie et al., 2001) is a sequential mining algorithm which uses the divide and conquer technique and was proved to be better than all other sequential pattern mining algorithms (Pie et
An analysis of the current and future of frequent pattern mining techniques is clearly explained by Han et al. (2007). Concept extraction from unstructured documents (discharge notes) was done using Apriori rule mining algorithm by Basak (2011). Waikato Environment for Knowledge Analysis (WEKA) tool was used to do the preprocessing and extraction of the concepts. An interface was used to give inputs to the tool.

Based on the important mining algorithms discussed above and the conclusions inferred so far, a decision was taken to use the FP-growth association rule mining algorithm and PrefixSpan sequential pattern mining algorithm to mine the association rules from the documents. Concepts were extracted using the existing BMI approach and proposed FP-growth and PrefixSpan algorithm and compared based on the recall and precision measures.

2.4 CLASSIFICATION OF SEMANTIC SIMILARITY MEASURES

There are different kinds of similarity measures proposed for evaluating single ontologies and multiple ontologies. The following sections discuss the inter concept similarity measures and intra concept similarity measures.

2.4.1 SIMILARITY MEASURES FOR SINGLE ONTOLOGY

Different approaches for measuring semantic similarity within the concepts in a single ontology are broadly classified into Edge counting approach, Information Content approach, hybrid based approach and feature based approach. They are given a brief introduction in this section.

Edge counting calculates similarity by counting the number of nodes between two concepts in terms of path length of the nodes and shortest path between the two concepts in the ontology is identified. The different path length based measures are the Rada measure, Hirst and St.Onge measure, Bulskov measure, Ehrig measure etc. Rada et al. (1989) calculated the similarity measure by calculating the number of minimal arcs which separate the two concepts in an ontology. This measure is based
on the Quillian spreading activation theory (Quillian, 1968) where the conceptual distance is used to quantify the similarity between concepts. Smaller the value, greater is the similarity between the concepts. The experiments were conducted using Medical Subject Headings (MeSH) ontology. Another method to calculate the path length between two concepts was proposed by Hirst and St.Onge (1998). They classified the semantic relations in WordNet Ontology into strong, extra strong and medium relations. This classification is done based on the number of paths available between the two concepts. The path length is permitted only if there are less than 5 links.

A few edge counting methods the depth of the ontology is taken into account by calculating the depth from the root node to the target node. The methods by Wu and Palmer (1994), Leacock and Chodorow (1998) and Sussana (1993) are measures calculated based on the depth of the concepts. Sussana has averaged the inverse weights (the edges leaving out and edges coming in from a particular edge) scaled by the depth of the taxonomy. Wu and Palmer consider the length between the concepts and the depth of the whole hierarchy to calculate the measure. Leacock and Chodorow (1998) proposed a measure that calculates similarity based on IS-A hierarchy for nouns in WordNet.

Information content methods calculate the similarity based on what the contents have in common. They are mainly corpus dependent. Analysis of large corpus of data is time consuming (Lin D., 1998; Resnik P., 1995). Resnik has defined his similarity measure based on the common amount of information shared between the two concepts. This work was extended by Lin by stating that the similarity of two concepts is the ratio of the common information shared to the information required to describe the concepts.

Hybrid methods calculate the similarity by combining any two measures discussed above. Few hybrid measures discussed here are that of Jiang and Conrath, Li measure and Zuber and Faltings measure. Jiang and Conrath (1997) combined the edge count and information content measure to find the similarity measure. It is calculated as the sum of individual nodes using the shortest path and by assigning weight to each edge based on the corpus statistics. The shortest path and local density of words were used to compute the similarity measure proposed by Li, (2003). Apriori
score between the two concepts was calculated by Zuber and Faltings, (2007). This score was transformed into a distance measure. This measure is normalized based on the depth of taxonomy created. It was tested using the WordNet ontology and Gene ontology.

Feature based method for calculating similarity measures is proposed by Tversky (1977). The similarity between the two terms is calculated as a function of their properties in feature based methods. Common features tend to increase the similarity between the two concepts and non-common features tend to decrease it. The Tversky’s approach was extended by Pirro (2009) by computing the similarity measure based on concept relations. The experiments were conducted using the WordNet and MeSH ontologies.

Other similarity measures that do not fit exactly into the broad classification of measures are discussed here. Pedresen et al. (2007) have proposed a measure to find the similarity between concepts in SNOMED-DT biomedical ontology. It is a combination of different measures. He also discussed how similarity measures and relatedness can be applied in different Natural Language Processing (NLP) applications (Pedresen, 2013). Al-Mubaid and Nguyen, (2006) have proposed a cluster based approach to measure the similarity among biomedical concepts. It used MEDLINE as a Corpus and MeSH ontology. Nguyen and Al-Mubaid, (2006) have proposed an ontology based approach to measure the similarities between biomedical concepts. Ontology based similarity measures have been of great importance in disambiguating text in biomedical domain (Bridget and Pedersen, 2013). Mabotuwana et al. (2013) have enhanced the Cosine similarity measure to compare the radiology reports. Context depended similarity measures were created to classify the clinical data (Garlaa,V.N., and Cynthia Brandt, 2012). The next section discusses the several measures available for finding similarities between multiple ontologies.

2.4.2. SIMILARITY MEASURES FOR MULTIPLE ONTOLOGIES

The similarity measures discussed above were used to measure concepts within a single ontology or taxonomy. With the exponential growth of resources, there is an impending need to compute measures that will help comparison between more
than one ontology or taxonomy (David and Montserrat, 2013). As of now, there are only a few measures available for cross ontologies. They are the path length measure (Al-Mubaid and Nguyen, 2009) and feature based measures (Rodriguez et al. measure, 2003; X-Similarity measure, 2006).

2.4.2.1 AL-MUBAID AND NGUYEN MEASURES

The path length based measure proposed by Al-Mubaid and Nguyen (2009) uses an ontology based approach to measure the similarities between biomedical concepts and also measured similarity using multiple sources. The experiments were done using Medical Subject Headings (MeSH) and SNOMED-CT ontologies defined in the Unified Medical Language System (UMLS) framework. Out of the two ontologies, the ontology with a higher density of concepts is considered as the primary ontology and the other ontology with lesser number of concepts is considered as the secondary ontology. Their measure helped in finding similarities between concepts in primary ontology, concepts belonging to primary and secondary ontology and concepts within secondary ontology. Here a brief discussion about cross ontology similarity measure is given. First the least common subsuming (LCS) concept is calculated using the formulae

\[
LCS(C_1,C_2) = LCS(C_1, bridge) 
\]

where \(C_1\) belongs to primary ontology and \(C_2\) belongs to the secondary ontology. \(bridge\) refers to the bridge nodes. Common nodes of both the concepts are merged together and they are termed as the bridge nodes.

The path length of the two concepts is calculated by adding the path length of each concept node to the bridge node.

\[
\text{pathlength}(C_1,C_2) = d_1 + d_2 - 1 
\]

where \(d_1\) is the shortest path between the concept \(C_1\) and bridge node and \(d_2\) is the shortest path between the concept \(C_2\) and bridge node. The common specificity is defined as
where $D_1$ and $D_2$ are the depth of the primary and secondary ontologies.

Finally the semantic distance between the concepts is given by the equation

$$SemDistance(C1,C2) = \log(pathlength_i - 1)^{\alpha} \ast (Commonspec_i)^{\beta} + K$$

In hybrid similarity measures, the term similarity is computed by matching synonyms, term neighborhoods and term features. (Rodriguez and Egenhofer, 2003; Euripides et al., 2006). Term features are further distinguished into parts, functions and attributes and are matched similar to Tversky’s method. The two hybrid similarity measures discussed here are the Rodriguez and X-similarity measures.

2.4.2.2 RODRIGUEZ AND EGENHOFER SIMILARITY MEASURES

Rodriguez, M.A. and Egenhofer, M.J. (2003) consider three parameters for the processing of the similarity between the concepts. The parameters are synsets, features and neighborhoods. The synsets are referred to as the weighted sum of synonym sets, features are referred to as the part of relations between the concepts and neighborhoods are referred to as the surrounding concepts of the selected concept.

In this method, the ontologies mainly used are the WordNet and MeSH. The method defined a similarity measure as follows:

$$Sim(a,b) = w.S_s(a,b) + u.S_f(a,b) + v.S_n(a,b)$$

for $w, u, v \geq 0$ and $w + u + v = 1.0$

The functions $S_s, S_f$ and $S_n$ are the similarity measures calculated among the synsets, features and term neighborhoods of the ontologies. $w, u$ and $v$ are the respective weights of the similarity of each component. The formulae for computing $S(a,b)$ for synsets, features and neighborhoods is given in the following equation.

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha (a,b)|A \setminus B| + (1-\alpha (a,b))|B \setminus A|}$$
where A, B denote synsets, features or neighborhoods of terms a, b and $A \setminus B$ denotes the set of terms in A but not in B (the reverse for $B \setminus A$). The above similarity measure is derived from the Tversky’s Model. Parameter $\alpha(a, b)$ is computed as a function of the depths of the terms $a$ and $b$ in their taxonomy.

$$\alpha(a, b) = \begin{cases} \frac{\text{depth}(a)}{\text{depth}(a) + \text{depth}(b)}, & \text{depth}(a) \leq \text{depth}(b) \\ 1 - \frac{\text{depth}(a)}{\text{depth}(a) + \text{depth}(b)}, & \text{depth}(a) > \text{depth}(b) \end{cases}$$

If the value of $\alpha$ satisfies $0 \leq \alpha \leq 0.5$, then both the terms are in the same depth or hierarchy and the terms are dissimilar. If the value of $\alpha$ and $(1-\alpha)$ are less than 1, then it means that there exist a similarity between the terms $a$ and $b$.

### 2.4.2.3 X-SIMILARITY MEASURE

X-similarity relies on matching between synsets and term description sets. The term description sets contain words extracted by parsing term definitions (Euripides, 2006). Two terms are similar if their synsets or description sets are lexically similar. From the above equation, a plain set similarity equation is deduced.

$$S(a, b) = \frac{|A \cap B|}{|A \cup B|}$$

where $A$ and $B$ represent the synsets or term descriptions, because not all terms are connected to neighborhoods with the same relationship type. A detailed form of expressions as per the X-similarity method to compute similarity between neighborhoods can be represented as,

$$S_n(a, b) = \max_i \frac{|A_i \cap B_i|}{|A_i \cup B_i|}$$

where $i$ denotes the relationship type. $S_n$ denotes the similarity between neighborhoods and $S_d$ denotes the similarity between term description sets. The above equation is used to compute the values of $S_d$ and $S_n$. The whole idea discussed above is concluded in the following equation.
This equation is used for finding the concept similarity between the two concepts in different ontologies.

The aim of our proposed work is to compare the existing hybrid similarity measures and derive an enhanced measure which outperforms the existing methods by producing similarity measures more closer human scores published by Intelligent System Laboratory, Technical University of Crete.

2.5 USE OF SEMANTIC TECHNOLOGIES FOR E-LEARNING

The Semantic Web can be seen as an extension of World Wide Web (WWW) where the information exists in a machine understandable format which facilitates the efficient exchange of information between computers and people (Fensel and Musen, 2001 and James handler et al. 2002). Leading Research, Educational and industrial institutions are having discussions about incorporating the Semantic Web technologies to enhance the storage and retrieval of related information effectively. E-learning (Thyagharajan and Nayak, 2007) is a domain which may profit from this new Web technology (Dutta, 2006). It is not just giving easy access to learning resources, anytime, anywhere, via a repository of learning resources, but it also includes features like customizing the learning schedule and giving materials based on the knowledge of the learner and collaboration among learners and between learners and instructors (Barker, 2000 and Durcker, 2000). E-learning, or improved computer supported learning, focuses on the individual's acquisition (or rather construction) of new knowledge and the technological means to support this construction process (Edward et al. 2007).

The learning requirements of a learner can be improved through proper guidance. A tutor helps to organize the learning resources based on the context of the learner. The resources from text books are transformed to computer based lessons. All the resources are related to one another based on the context of the resources (Andreas Schmidt, 2005). The constraints in the traditional learning systems were got over by
the current e-learning systems. Current e-learning systems are empowered with the enhancements of technology and are designed to meet the learner’s need in a more sophisticated way. Such systems are called adaptive e-learning systems. In adaptive systems, the learning materials are organized based on the context and the knowledge level of the learners.

Jelena Jovanovic et al. (2007) have shown the method of using Semantic Web technologies to get better state-of-the-art in online learning environments and overcome the gap between students on the one hand, and tutors on the other. The ontological framework on hand helps in formalize learning object context as a multifarious interplay of different learning-related elements and points how they could use semantic annotation to correlate diverse learning artifacts. On top of the framework, they have provided several feedback channels for educators to develop the delivery of future Web-based courses. Ig Ibert Bittencourt et al. (2009) incorporated the semantic web techniques into AIED (Artificial Intelligence in Education) systems to enhance the educational systems. They have developed a computational model that creates a semantic structure of Web pages, deploy software agents to navigate through the Web documents. This helps in understanding the content inside the Web applications and in performing difficult tasks. Ontologies are used to define and relate the data in the documents. Four ontologies used here were: the domain ontology, student model ontology, pedagogic model ontology, interaction ontology. Semantic web services are used to help the agents in providing services. This computational model is illustrated through the development of an intelligent tutoring system.

The expressive power of semantic tools and technologies to describe learning contents, identify the level of learners and the knowledge they have and match them intelligently with the required learning materials, can improve the standards of learning through Web (Thanassis Tiropanis et al. 2009). They have classified semantic technologies into two types, namely hard and soft. Hard semantic technologies help us to express the meaning of resources and establish a relationship between them, and draw conclusions based on the meanings. Examples are RDF and Friend Of A Friend (FOAF). Soft semantic technologies help us to express the meaning of resources in the way a human being can understand. Examples include...
tagging tools and topic maps. A survey was conducted and the semantic technologies for higher education were identified and categorized as follows. They were collaborative authoring and annotation tools, searching and matching tools, repositories and virtual learning environments, infrastructural tools and services. Later, Web 2.0 technologies enabled annotation of large volumes of data and introduced the concept of linked data which had a more hopeful future and benefit for higher education and learning contexts (Thanassis Tiropanis et al. 2012; Colin et al. 2012).

The need for an ontology based approach for adaptive e-learning was suggested by Sohaib Ahmed et al. (2010) which will enable reusability of contents for any domain. Using ontology based method will help in providing the learning materials as per the needs of the individual learner. To assess the possibility of the proposed approach, they have built a web-based learning environment for first-aid learners. Reusability of the contents may guide learners during the learning process and also may be supportive for building knowledge domains within a Learning Management Systems (LMS). To enhance the adaptability, students were allowed to build the personal domain ontologies and thesauri based on the text books, lectures and manuals. These ontologies are compared with the ones built by the tutors to analyse the mistakes and notification about it is sent to improve the course materials of the students. A prototype was developed by Gladun, A. and. Rogushina, J. (2010) to control student skills.

A semantic e-learning environment was developed by Emina Junuz (2009) which uses two types of ontologies. One ontology is for personalization of learning based on preferences and the other ontology is for physical structure learning objects. Learning object (LO) is an entity, which can be reused again and again during the learning process. Some of the standard learning objects are IEEE LOM, Dublin Core, IMS, and SCORM. LO’s can be created at local and global level, where ontology gives LO’s there meaning and relationships. This helps the learner in finding the learning materials and the teacher for creating the course material.

Semantic web and domain ontologies serve as a backbone for e-learning by facilitating information sharing and exchange (Chakkrit and Michael, 2007). Xing Jiang and Ah-Hweeta (2009) developed a system for personalized semantic search.
An ontology based user model called user ontology is designed, which consists of all concepts (nodes) and relations of domain ontology together with a value indicating the user’s personalized interest. The spreading activation theory is used for learning the ontology and inferencing it. For example, a node is assigned the activation value of 1.0. Depending upon the relationship with the neighbors of the node, the activation values will be assigned to each node. The activation then propagates through the entire semantic network. A Bayesian solution is used to compute the learning interest of the user and based on the cosine similarity measure, the documents in the initial list are re-ranked. This is applied to the dataset of Google directory and applied for document retrieval.

The above discussed research works formulate different ways of knowledge extraction from predefined ontology, automating an ontology and the use of ontologies in E-learning. All these approaches end up in designing a framework for deriving an ontology with the help of human intervention in between each phase of ontology building. Our proposed research work is to design a framework to automate an ontology from a given set of documents. The resultant ontology is in the form of an OWL file. This comparison is done using an appropriate cross ontology similarity measure. Different cross ontology measures are discussed in section 2.4. Based on this, a new concept similarity measure is designed which is an enhanced version of the hybrid cross ontology measure.

One of the important fields that will benefit by this proposed research work is E-learning. Section 2.5 gives an idea regarding the impact of semantic web technologies in the field of e-learning. The automated ontology is compared with the domain expert’s ontology generated manually. This comparison is done based on the new derived similarity measure. This measure can be embedded into any e-learning system for more efficient retrieval of learning documents.