IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 2, Issue 4 (July-Aug. 2012), PP 38-48
www.iosrjournals.org

Concept Based Information Retrieval from Text Documents

Ms.D.Subarani, Assistant Professor,
Dept. Of Computer Sciences,SLN College Of Sciences, PalluruAffiliated to S.V University, Tirupathi. INDIA.

Abstract: *This research is intended to develop a concept based information retrieval system for text documents in two phases: Therefore, this idea motivated us to develop a concept based information retrieval system for text documents. This system tries to provide additional semantics as conceptually related words with the help of glosses to the query words and keywords in the documents by disambiguating their meanings. Here, various senses provided by WSD algorithm have been used as semantics for indexing the documents to aid the information retrieval system. Later, this research has also been motivated to do ontology based information retrieval from Tamil text documents which improve the retrieval performance in a better way due to the incorporation of domain semantics.*

Here, the performance of IR has been improved by including more indexing information about the documents such as associated meaning with the words. The Word Sense Disambiguation is the process of finding correct senses of a word, among other senses associated with the words. The introduction of semantics in word level to improve Word Senses Disambiguation has been considered in this thesis specifically to improve the accuracy of WSD and thus in turn to improve the IR performance. In this work, the glosses of the indexed words in WordNet are utilized as conceptual information, which acts as an additional semantics for WSD. This concept based WSD, has been used in semantic chaining to cluster documents, which is used for IR performance.

Keywords: *Word Senses Disambiguation (WSD), Information Retrieval (IR), Ontology.*

I. Introduction

Information retrieval can be defined as a system for representing, indexing (organizing), searching (retrieving) and recollecting (delivering) documents. The goal of IR is to provide users with documents that will satisfy their information need. In other words, the query given by the user has to match with the available index of the document.

(A)Document Databases

The purpose of a database is to provide a wrapper through which one can retrieve information, which is relevant to a given query. Depending on the nature of the data to be accessed different techniques must be used for creating indexes, formulating queries and retrieving records. Maintaining and accessing a database of full-text documents is a more challenging problem.

A realistic expectation for today's text document database systems is not to provide answers, but rather to provide mechanisms to retrieve documents, which are most relevant to the formulated query, in ranked order. The text retrieval community has made significant progress in Information Retrieval (IR).

(B) Information Retrieval

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information need, for example, search strings in web search engines. In information retrieval a query does not uniquely identify a single document in the collection.

Instead, several documents may match the query, perhaps with different degrees of relevancy. Depending on the application the document may be text documents, images or videos. Often the documents themselves are not stored directly in the IR system, but are instead represented by document surrogates. Most IR systems compute a numeric score on how well every document in the collection matches the query, and rank the documents according to this value. The top ranked documents are then shown to the user.

The users normally interact with the system through a query, which is essentially a string of words that characterizes the information that the user seeks. The user of a retrieval system has to translate his/her information need into a query in the language provided by the system. This normally implies specifying a set of words, which convey the semantics of the information need.

The user query is then given to the IR system to retrieve information, which might be useful or relevant to the user by matching the query with the index. The simplest index for the collection of documents is frequently occurring words known as bag of words. The information retrieved is ranked according to the strength of relevancy between the document and the query. The ranking of the documents can be done by

various similarity matching methodologies. User relevance feedback can be used to reformulate the query for better retrieval. Query expansion or translation can be done to retrieve more number of documents.The information retrieved by IR systems needs to be relevant with respect to user query. Relevance depends not only on the query and the collection but also on the context, i.e., the user's personal needs, preferences, knowledge, expertise language etc. In addition, relevance of a document may depend on the order of retrieval. Hence it is necessary to evolve a methodology to objectively evaluate the performance of an IR system.

Six major measurable quantities including the coverage of collection, the time lag, the form of presentation, the effort, the recall, and precision were proposed. Of these, recall and precision are commonly used parameters utilized to measure the effectiveness of the retrieval system. Recall is defined as the ratio of relevant items retrieved to all relevant items or the probability given that an item is relevant that will be retrieved. Precision is defined as the ratio of relevance items retrieved to all items retrieved, or the probability given that an item is retrieved it will be relevant. These parameters are essentially a measure of the ability of the system to retrieve relevant documents while at the same time avoiding non-relevant one. It is normally considered that recall and precision are sufficient to measure retrieval effectiveness.

Major models such as Boolean model, Statistical models and Linguistics and Knowledge models have been developed to represent and index information for retrieval. Boolean model is an exact match model whereas the other models are best match models. The indexes derived for the documents have to be stored effectively for retrieval. Storage structures specifically define the logical organization of this information. Earlier document retrieval systems normally adopted serial file organization whereas the currently inverted files are used.

In the context of web search, a critical goal of successful IR is to identify which web pages are of high quality and relevance to a user's query. Current IR systems place most of the burden on the users, relying on them to identify sources likely to contain relevant information, compose an appropriate query, and sift through retrieved pages to extract relevant information. Clearly, as web pages continue to grow, it will become impractical for users to perform these tasks for all but the simplest requests.

(C) Web Search

IR encompasses many types of information access. Web search is an important part of this spectrum of information systems. Web search now represents a significant portion of Web activity. There are three forms of searching the Web. They are:

1. Using search engines that index a portion of the web documents as a full text database.
2. Using Web directories which classify web documents by subject.
3. Searching the web by exploiting its hyperlink structure (sharedcontext).

General web search is performed predominantly through text queries to search engines. Because of the enormous size of web, text alone is usually not selective enough to limit the number of query results to a manageable size. Search engines for the web are one of the most publicly visible realizations of information retrieval technology. Search engines are critically important to help users find relevant information on the World Wide Web. For some search tasks (e.g., home page finding), systems such as Google (2009) provide highly accurate results. However, the Web contains more than just home pages and web users are interested in more than just finding single pages. Therefore, significant challenges remain for improving general Web search.

In order to best serve the needs of users, a search engine must find and filter the most relevant information matching a user's query, and then present that information in a manner that makes it readily palatable to the user. Moreover, the task of IR and presentation must be done in a scalable fashion to serve the hundreds of millions of user queries that are issued every day to a popular web search engines such as Google Danny Sullivan of Search Engine Watch (Searches Per Day 2009) estimates that eight major search services serve up over 625 million search requests per day.

(D)Areas Related to IR (Context of IR)

IR systems form a component of the larger issue of information analysis and search. Text mining, of which IR is a part, is defined as knowledge discovery in large text collections. Its aim is to use knowledge representation techniques for representing the domain of interest and to extract relevant data from the document. This will allow users to extract text containing the salient concepts. In other words, text mining is about looking for regularities, patterns or trends in natural language text, and usually is about analyzing text for particular purposes. Complementary to IR, is text categorization which can help to overcome the information overload problem. Categorization systems automate the assignment of documents to a hierarchical organization of topics but require training and prior knowledge of the topics in a corpus. A practical solution is to discover and approximate these topic hierarchies using unsupervised clustering methods. The clustering is the process of grouping together similar documents to improve IR performance. One technique involves feature extraction, which maps each document to a point in high dimensional space, and the clustering algorithm automatically

groups the points into a hierarchy of clusters. Text clustering is used to provide a overview of content in a large document collection, identify hidden structures between group of objects, find outstanding documents within a collection, detect duplicate documents in an archive.

The dramatic growth in the number and size of on-line textual information sources has led to an increasing research interest in the information extraction problem (IE). IR retrieves relevant documents from collections. Once this is done, IE automatically extracts information from the searched documents according to a predefined template. Hence, the two techniques are complementary, and used in combination they can provide powerful tools for text processing. The IR and IE differ not only in aims but also in the techniques usually deployed. Most of the work in IE has emerged from research on rule-based systems in computational linguistics and natural language processing while information, probability theory and statistics have influenced IR.

II. Another related area is text summarization. Extraction of pieces of an original text on a statistical basis or with heuristic methods and putting it together to form a new shorter text with the same is one method of text summarization. There are three steps to perform text summarization. First to understand the topic of a text, so called topic identification, secondly the interpretation of the text and finally the generation of the text. Text extraction techniques basically give scores to each sentence depending on the importance of each sentence and when creating the summary the most significant sentences are retained. The scores can be based on high-frequent open word class words, bold or numerical text, proper nouns, citations, position in text etc. Since text retrieval is going to be performed in Tamil text documents, an overview of the Tamil language has been presented in the next section.

(E) Tamil Language

Tamil is a South Indian language spoken widely in Tamil Nadu in India. Tamil has the longest unbroken literary tradition amongst the Dravidian languages. Tamil is inherited from Brahmi script. Tamil has 12 vowels and 18 consonants. These are combined with each other to yield 216 composite characters and 1 special character (aayatha ezhuthu) counting to a total of (12+18+216+1) 247 characters.

Vowels

Vowels in Tamil are otherwise called UyirEzhuthu and are of two types short (Kuril) and long (Nedil).

Consonants

Consonants are classified into three classes with 6 in each class and are called Vallinam, Idaiyinam, and Mellinam.

(F) Motivation of this Research

As the information to be handled is a free form of natural language text, Natural language processing techniques represents a challenging task. A limited amount of syntactic and semantic analysis has to be done to extract the units from the NLP text. The units, words, senses, word co-occurrence, syntactic word grouping and semantic relations between constituents are needed to represent the documents in richer form to get better retrieval performance. The indexing of the document needs semantics to change the word-based approach to sense based approach for effective retrieval. The synonymy and polysemy effect can be solved by this semantic indexing. The sense based component of the retrieval system eliminates the possibility of retrieving documents that are obtained due to the presence of polysemes of the keywords. The word sense disambiguation algorithm is needed for semantic indexing to get the correct sense of the indexed words.

The semantic representation of sentence constitutes make the indexing to move from two dimension to third dimension with thematic role relation. The thematic role coupling with senses makes the representation more informative than only with words. The content level extraction from the document makes the representation as logical which is provided with inference. The inference mechanism in logic models handles the incomplete and inconsistent queries. The query also interpreted efficiently with semantic based approach. As a result, this research has been motivated to develop an information retrieval for Tamil text documents using concepts of the text rather than the keywords.

(G) Contributions of this Research

As guided by the above motivations, this research intends to develop a concept (semantic) based information retrieval from Tamil text documents. Semantics has been introduced at various linguistic levels, word level, sentence level and document content extraction level and at various stages of Information Retrieval such as query and document representation, and indexing, in this thesis to improve the information retrieval from text documents. However, it is to be emphasized that any attempt to bring in semantics needs to balance the amount of complex natural language processing required, with the increase in retrieval performance.

As a result of the above motivations, following contributions arise from this research:

1. Word Sense Disambiguation has been performed over the words which improves the information retrieval performance semantically over Tamil text documents.
2. Domain ontology has been created with knowledge bases to support semantic search in Tamil document repositories which also improves the retrieval performance conceptually.

II. Materials And Methods

Concept based Information retrieval system from text documents has been developed successfully by using JAVA as an implementation platform and a combination of Word Net & Ontology has been used a lexical resource as a material for this research.

(A) Wordnet

The reason for choosing the WordNet as the lexical resource over any other online thesaurus is that the WordNet not only provides the user with the meaning of a word but in addition provides semantic relations such as synonyms, hypernym, hyponyms and antonyms of that word. WordNet divides words into synonym sets or *synsets*, groups of words that are synonyms of one another. These synsets are then connected by a number of different relations. A particular word may appear in several synsets, depending on how many senses it has[ ]. These synsets are then inter-connected as a net of synsets by links on a number of different relations such as the following:

- IS-A relation (Hyponym). Eg. Apple *is a* fruit.
- INCLUDES relation (Hypernym). Eg. Fruits *include* apple.
- ANTONYM relation. Eg. Boy is an *antonym* of girl.

Relevant pointers to form a linked list of synsets connect all these synsets.

The WordNet lexical database is supported by an application-programming interface, which consists of certain data structures and primitive utility functions to access the database. There are separate WordNet APIs for Windows platform and Linux. Functions within the WordNet database, which are essential to the work described in this thesis, are search and morph functions. Four types of search, synonyms, antonyms, hypernyms and hyponym search, can be performed in the WordNet to get the corresponding semantically related words. The WordNet Structure for Noun is shown in fig .1

Hypernym synset for W1

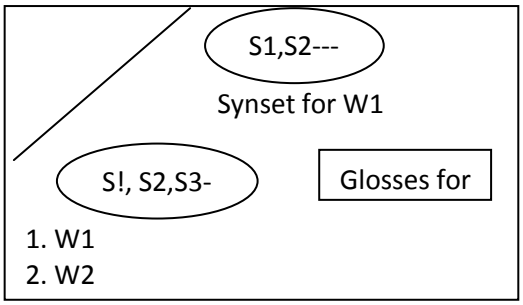

Figure 1 The WordNet Structure for Noun

A word may have many senses. If the synset of a word for a particular sense is needed, the sense number can be passed to this function as *Sense*. Words are often associated with multiple senses i.e. display polysemy. When a word is given to the WordNet a corresponding set of synsets containing all senses of the word is obtained. The disambiguation process aims at choosing the correct sense of the word.

The POS tag and the root of the content words of the documents are obtained using the WordNet. The output of the preprocessing phase which removes all the unwanted words like prepositions, articles etc, is thus a set of root words and their associated unambiguous POS tag

i.e .Stem(root)| POS..

(B) Ontology

Ontology is a specification of a conceptualization. The word "ontology" seems to generate a lot of controversy in discussions about AI. It has a long history in philosophy, in which it refers to the subject of existence. It is also often confused with epistemology, which is about knowledge and knowing. The term "ontology" can be defined as an explicit specification of conceptualization. Ontologies capture the structure of the domain, i.e. conceptualization. This includes the model of the domain with possible restrictions. The conceptualization describes knowledge about the domain, not about the particular state of affairs in the domain. In other words, the conceptualization is not changing, or is changing very rarely. Ontology is then specification of this conceptualization - the conceptualization is specified by using particular modeling language and

particular terms. Formal specification is required in order to be able to process ontologies and operate on ontologies automatically.

Ontology describes a domain, while a knowledge base (based on ontology) describes particular state of affairs. Each knowledge based system or agent has its own knowledge base, and only what can be expressed using ontology can be stored and used in the knowledge base. When an agent wants to communicate to another agent, he uses the constructs from some ontology. In order to understand in communication, ontologies must be shared between agents.

Explicit specification of conceptualization means that an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set of concept definitions, but more general.

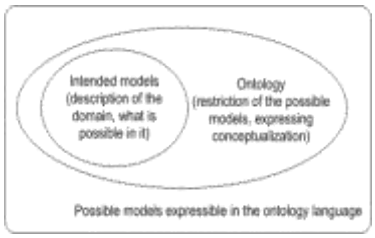

Figure 2 Ontology expressing intended models for description of the domain

A conceptualization can be defined as an intensional semantic structure that encodes implicit knowledge constraining the structure of a piece of a domain. Ontology is a (partial) specification of this structure, i.e., it is usually a logical theory that expresses the conceptualization explicitly in some language. Conceptualization is language independent, while ontology is language dependent.

In this sense, ontology is important for the purpose of enabling knowledge sharing and reuse. Ontology is in this context a specification used for making ontological commitments. Practically, an ontological commitment is an agreement to use a vocabulary (i.e., ask queries and make assertions) in a way that is consistent (but not complete) with respect to the theory specified by an ontology. Agents then commit to ontologies and ontologies are designed so that the knowledge can be shared among these agents.

The representation of a body of knowledge (knowledge base) is based on the specification of conceptualization. A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system or agent is committed to some conceptualization, explicitly or implicitly. For these systems, what "exists" is what can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects and the describable relationships among them are reflected in the representational vocabulary with which a knowledge-based program represents knowledge.

Thus, in the context of AI, ontology of a program can be described by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g. classes, relations, functions, or other objects) with descriptions of what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally it can be said that an ontology is a statement of a logical theory.

The backbone of ontology is often taxonomy. Taxonomy is a classification of things in a hierarchical form. It is usually a tree or a lattice that express subsumption relation - i.e., A subsumes B meaning that everything that is in A is also in B. An example is classification of living organisms. The taxonomy usually restricts the intended usage of classes - where classes are subsets of the set of all possible individuals in the domain. Taxonomy of properties can be defined as well.

However, ontology need not to be limited to taxonomic hierarchies of classes and need not to be limited to definitions that only introduce terminology and do not add any knowledge about the world. To specify a conceptualization, axioms that constrain the possible interpretations for the defined terms may be also needed. Pragmatically, ontology defines the vocabulary with which queries and assertions are exchanged among agents. The ontological commitment is then a guarantee of consistency for communications.

Ontologies, if they are to be used for automatic processing in computers, need to be specified formally. There are several languages that are used for expressing ontologies. The formality of the description of ontologies is summarized in the figure 3. On the right end of the scale there is a catalog of terms used for expressing knowledge or information. These terms may have no description at all, and they are understood only because they are chosen from the natural language, and their meaning can be only estimated. The description of each term in natural language is better, especially if there are also relations expressed between the terms, such as is-a, part-of, related-to, etc. However, until this description is in natural language, which is not formally defined, we usually do not call such specification ontology.

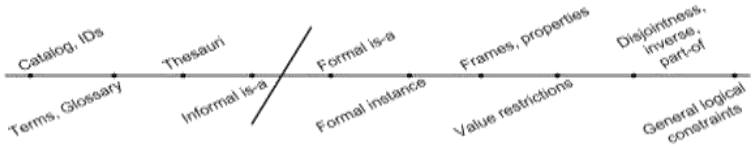

Figure 3 Formal Descriptions of Ontologies

In the context of knowledge sharing, we will use the term ontology to mean a specification of a conceptualization. That is, ontology is a description of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy.

Designing ontologies for the purpose of enabling knowledge sharing and reuse is the primary functionality of ontology. In that context, an ontology is a specification used for making ontological commitments. The formal definition of ontological commitment is given below. For pragmatic reasons, we choose to write an ontology as a set of definitions of formal vocabulary. Although this isn't the only way to specify a conceptualization, it has some nice properties for knowledge sharing among AI software (e.g., semantics independent of reader and context). Practically, an ontological commitment is an agreement to use a vocabulary (i.e., ask queries and make assertions) in a way that is consistent (but not complete) with respect to the theory specified by an ontology. We build agents that commit to ontologies.

III. Concept Based Word Sense Disambiguation For Information Retrieval

(A) Introduction

The words that appear in documents have many semantic senses depending upon the local context. Words need to be disambiguated and the context in which they occur in a particular document needs to be determined. In general terms, WSD (Word Sense Disambiguation) involves the association of a particular sense (meaning) to a word in a document, among other senses that can be potentially attributed to the word. The disambiguated words are essential in applications like information retrieval, information extraction, text summarization, and all tasks in a text-mining framework. Using a more efficient WSD process before the actual retrieval process minimizes irrelevant information retrieval.

In traditional word based approach, the irrelevant information is retrieved because of non provision of the correct sense of the word in the searching process and this can be rectified by the sense based approach. The sense provided by WSD process is added as semantic information in association with the given word in semantic indexing oriented sense based IR. The different WSD algorithms differ in the correctness by which they chose the sense and the type of linguistic information they use to do so. In evaluating the different WSD algorithms in the context of IR it is also important to note that associating wrong senses and discarding correct senses will adversely affect retrieval performance.

Lexical information of the word in the form of lexical relations like synonym, hypernym, hyponym etc taken from the WordNet, has been used for WSD by many researchers. In the work described in this thesis conceptually related words which are not directly available in the WordNet but derived from the content words of the glosses in the WordNet provides additional semantics to improve the accuracy of the WSD process.

(B) Word Sense Disambiguation

Word sense disambiguation is an open problem in Natural Language Processing (NLP). Selecting the most appropriate sense of an ambiguous word in a sentence is a central problem in NLP. For representing or understanding the NLP sentence, the correct senses of the content words of the sentences are necessary. The information retrieval is a task where WSD enables to obtain better performance.

While there are a variety of different ways the retrieval can be accomplished, most systems treat the query words as a pattern to be matched by documents represented as words. Unfortunately, the effectiveness of these word-matching systems is decreased by both homographs and synonyms. Homographs decrease the accuracy of the retrieval systems by making texts about different concepts appear to match. Synonyms impair the system's ability to find all matching documents, since different words mask conceptual matches.

While polysemy is the immediate cause of the first problem, it indirectly contributes to the second problem as well by preventing effective use of thesauri. These considerations have motivated the investigation for a highly accurate word sense disambiguator. The information from a lexicon or knowledge bases are used in WSD for getting semantically related words which is essential in finding the correct senses of the word. The WordNet, LDOCE and Thesaurus are used as lexical knowledge sources. The semantically related words of a given word are essential to find the correct senses of the word. In this thesis, WordNet, the lexical resource is used for finding semantically related words.

(C) Sense Based Word Sense Disambiguation

The Sense based WSD algorithm used for IR is an iterative algorithm does not consider the domain oriented conceptual information associated with the documents. WordNet is used for detecting the correct sense of the word, however they do not use additional domain knowledge. The domain dependent conceptual information is taken from content words of the glosses of Wordnet and used for determining the correct sense of the word. The gloss of the Wordnet in general gives a description containing conceptually related words associated with the word under consideration. The concept based WSD is particularly useful when there is an ambiguity in determining the correct sense of a word. In such cases additional information in the form of association with other concepts obtained from the glosses of WordNet is used for the disambiguation process. This procedure has been also used to disambiguate the words in the query.

The concept based Word Sense Disambiguation (WSD) algorithm described in this section has been also used to find the correct sense of the words in lexical chains where the lexical chains are formed for finding the relatedness of documents. Then the documents are clustered which in turn improves the performance of IR. The semantic chains described in this thesis are formed with only the synsets corresponding to the correct sense of the bag of words of the document while the lexical chains are formed with all the synsets corresponding to the words. The synset weight vectors are formed with the semantic chains of the document. The linking between the synset weight vectors of the different documents is obtained in the same way as for lexical chains. However, the effectiveness of the process is increased due to the association of only the correct sense with the word. The use of Lexical resource for hierarchically clustering request for comments documents have also been studied.

WSD module performs a semi complete but precise disambiguation of the words in the documents. From the given set of synsets, the disambiguation process selects the correct sense of the word. When the POS tag and a sense number is given to the WordNet, the corresponding unique number called offset is obtained. The output for each word is now represented as Stem | POS | Offset where Stem is the Stemmed form of the word, POS is the Part of Speech and offset is the offset of the WordNet synset in which this word occurs. A unique identifier called the offset describes each synset in WordNet. This offset points to a set of synonymous words (synset). An offset indicates only one sense of a word and this offset has been chosen after disambiguation. Hence in the context of use of WordNet the offset gives the correct meaning of the word. The words with this kind of semantic information are useful for some applications in Text mining framework. The basic WSD algorithm described above has a major drawback in that it does not consider domain information.

(D)Concept Based WSD

Key words have often multiple functionalities (i.e. can have various parts of speech) or have several semantic senses. The disambiguation process relies on semantic information for identifying the meaning of the words and is based on WordNet senses. Mihalcea et al took only lexically related words for finding the correct sense of the words. The lexical relations like synonymy, hypernymy, hyponymy etc are taken from the lexical resource, WordNet. The conceptually related words also needed for detecting the correct senses of the words. In the approach described in this thesis, domain dependent conceptual information is taken from content words of the gloss associated with the word for finding its correct sense.

The concept based WSD algorithm modifies Mihalcea's work by including the conceptually related words and also considering Hypernym synsets. The conceptually related words are taken from the content words of the glosses. The glosses, which are description of words, are taken from the WordNet. The concept based WSD is based on the iterative WSD algorithm presented by Mihalcea and Moldovan. The semantic tagging is performed using the senses defined in WordNet. The various steps to identify the correct sense of a word are presented in this section.

The modified WSD algorithm is outlined below.

Identification of Monosemous words: The words having only one sense in WordNet are identified and then marked with sense #1. Here, there is no ambiguity and hence this procedure gives 100% accuracy.

Example: The words Exam, authorize etc have one sense defined in WordNet. Thus, it is a monosemous word and can be marked as sense #1.

A sense number of 0 is given to those words, which are not present in WordNet. In this case, indexing is done on the word instead of sense in Information Retrieval process.

Example: Proper nouns like India, Atal Bihari Vajpayee etc.

Determination of correct sense based noun context: For a word identified as noun, the noun context of each of its senses is determined. The Noun context is a list of nouns, which can occur within the context of a given sense i of the noun, N. For each sense of the word, the number of content words in the gloss and the words in the synset and Hypernym synsets are added to the list. Since gloss is the description of a word, it conveys the conceptual knowledge.

For each word j and for all sense i of word j, find

$NCsij = \{S_{sij} \cup HS_{sij} \cup G_{sij} \cup HG_{sij}\} \cap DC$

Where i= 1….m, j = 1….n,

n=number of words

m= Possible Senses of a word

DC – Document content words

NCsij – Noun Context of ith sense of a word j

HSsij – Hyponym Synset of ith sense of a word j

Ssij – Synset of ith sense of a word j

Gsij – Content words of the glosses of ith sense of a jth word and

HGsij - Content words of the glosses of Hypernym of ith sense of a jth word

Then the number of common words between this noun context and the original text, in which the noun N is found, is calculated. Applying this procedure to all the senses of the noun will provide an ordering over possible senses

Nij = Number of words in NCsij

For each word j, find i which is having maximum Nij. Then i denotes the sense of j. Then the sense i of the noun N that is in the top of this order is picked up as sense number.

The determination of Noun context is shown in figure 4

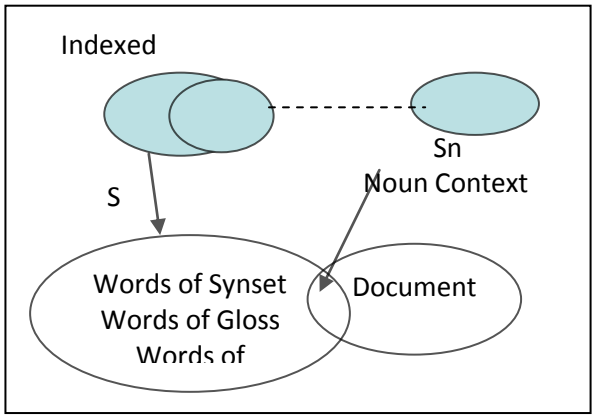

Figure 4 Determination of Noun contexts

Example: The word Hockey has two senses. The noun contexts of its senses are

hockey#1**{Play,Field,Opposing,Team,Sticks,Drive,Ball,Net,Opponent**}&

hockey#2{*Game,Played,Ice,rink,Goal,Stick,Opposing, Skaters,Knock,Puck,Flat,Opponent}.*

Based on the number of common words of the noun context with the original text, intended sense of keyword is concluded with the sense.

Determination of sense of words belonging to same synset as already disambiguated words:

Two words belonging to the same synset are said to be semantically connected at a distance of zero. The words, which fall in the same synset of the already disambiguated words, are found.

Example : Consider the two words ‘authorize’ and ‘clear’ both of which appear in the documents to be disambiguated. The verb, ‘authorize’ is a monosemous word, and thus it is disambiguated with step 1. One of the senses of the verb ‘clear’, namely sense # 4 appears in the same synset with ‘authorize’ #1, and thus clear is marked as sense #4. ‘Authorize’ of sense #1 and clear of sense #4 has a connection distance of 0.

Determination of sense of words belonging to same synset but not already disambiguated.

From the set of words, which are still not disambiguated, find the words, which are semantically connected among those that are yet to be disambiguated, with connection distance 0. Here all the senses of both words are considered in order to determine whether or not the connection distance between them is 0. One more level is taken by comparing with all the senses of the words in the hypernymy/hyponymy synset. This step helps to disambiguate words, which have no semantic connection to already disambiguated words.

Example:

For the two ambiguous words ‘jump’ and ‘leap’, this procedure tries to find two possible senses for these words, which are at a distance of 0, i.e. words belonging to the same synset or/and same hypernym/hyponym synsets. Sense #2 of Jump and leap have both synsets same. For jump and fly, in sense #2 both have same hypernym synsets.

Jump#2 jump,leap

Leap#2 leap,jump

Jump#2 leap

_> move
Fly#2 fly
_> move
The sense 2 is assigned to jump and leap.

Determination of senses of words belonging to same hypernymy/hyponomy synset already disambiguated:

The words, which are semantically connected to the already disambiguated words for which the connection distance is 1are found, i.e. they belong to a hypernymy/hyponomy relation. The words with sense i, which falls in the same hypernymy/hyponomy synset for the already disambiguated words are marked.
Example: The two nouns 'Subcommittee' and 'Committee'. The first is disambiguated with procedure 2 and it is marked as sense #1. Hypernymy relation semantically links the word 'Committee' with its sense #1 with the word 'Subcommittee'. Hence, the word 'Subcommittee' is marked with sense#1.

Default sense attachment: The words, which are present in WordNet and have yet to be disambiguated, are marked with sense #1The first sense of the word is the commonly occurring sense for a word. Hence the remaining words are assigned with sense #1.

The above steps are applied iteratively to identify a set of words, which can be disambiguated with high precision. In this algorithm the connection distance of one alone is considered. As connection distance increases i.e. one goes higher up the hypernymy chain the abstraction obscures. the special domain characteristics of the words of the document. The next section explains the concept based WSD algorithm, which uses the above-mentioned procedures iteratively for all the content words of the documents.

(E) Concept Based Wsd Algorithm

1. First the document is preprocessed. The Part of Speech tag is attached with each content word using WordNet. The stemming is done using the WordNet stemming algorithm.
2. Initialize the set of disambiguated words (SDW) with the empty set SDW={}. Initialize the set of ambiguous words (SAW) with the set formed by all the nouns and verbs in the document. Once the words are disambiguated, they are moved from SAW to SDW.
3. The monosemous words identified by step identification of monosemous words are removed from SAW and added to SDW.
4. Determination of correct sense based noun context will identify a set of nouns, which can be disambiguated based on their noun-contexts.
5. Determination of sense of words belonging to same synset as already disambiguated words tries to identify a synonymy relation between the words from SAW and SDW.
6. Determination of sense of words belonging to same synset but not already disambiguated is different from the previous one, as the synonymy relation is sought among words in SAW.
7. Determination of senses of words belonging to same hypernymy/hyponomy synset already disambiguated tries to identify words from SAW, which are linked at a distance of maximum 1 with the words from SDW.
8. Default sense attachment determines the remaining words, which could not be disambiguated but are present in the WordNet. Those words are marked with sense #1.

(F) Result of WSD Algorithm

The concept based WSD algorithm used in this work finds the specific sense in which the word occurs in a document. If there is ambiguity, the algorithm aims to pick out the most semantically correct sense by using additional domain dependent information for disambiguation. The difference in accuracy in the disambiguation process can be observed when concept information is used and when it is not.
Testing has been done for more than 500 documents from the sports domain. The number of words, which are disambiguated by a particular step and its accuracy, are calculated cumulatively without concept based and with concept based WSD algorithm.

IV. Semantic Based Information Retrieval From Tamil Documents

(A) Keyword Based Information Retrieval System

In this thesis, a system has been developed to retrieve information from digital library documents which is large collections of scanned documents (newspaper, books and journals) based on their conceptual information. Initially a keyword based information retrieval system has been developed as explained in section (i), which tries to retrieve the contents of the documents based on the occurrence of their keywords. Section (ii) discusses about the development of a semantic based information retrieval system which makes use of the

domain ontology to retrieve the conceptually relevant documents rather than keywords. Here Keyword based Search system has been implemented for Tamil documents in the context of banking domain. Steps involved in the keyword based search system have been explained below as indicated in figure 6.1.

- Documents Collection
- Pre-Processing of text documents.
  - Tokenizing and Special Characters Removal.
  - Stop word Removal.
- Document Annotation
  - Root word Extraction.
  - Verb and Noun identification.
  - Frequency count for nouns stored in DB.
  - Duplicate verbs are removed and stored in DB.
- Keyword based information Search

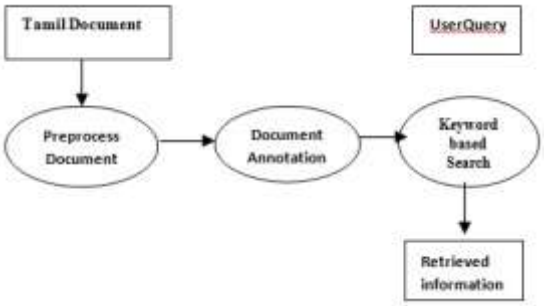

Figure 5 Keyword based Information Retrieval System

(i)Preprocessing of Tamil Documents

This module performs the preprocessing of Tamil documents which enables them for further processing. The preprocessing of Tamil Documents includes Tokenization, Special Character Removal and Stop word removal from the Tamil Documents as indicated in figure 6.

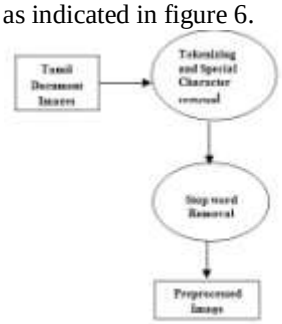

Figure 6 Preprocessing of Tamil Documents

(ii)Extracting Text from preprocessed documents

In this phase, text has been extracted from the preprocessed Tamil documents. From the preprocessed documents, root words are extracted after the removal of stop words. Later noun words are extracted from those terms as a result of Part of Speech Tagging. Frequency of the noun words collection has been taken followed by the removal of duplicate verbs. Later these terms are stored in the database to enable the retrieval process. This process has been indicated in the figure 7.

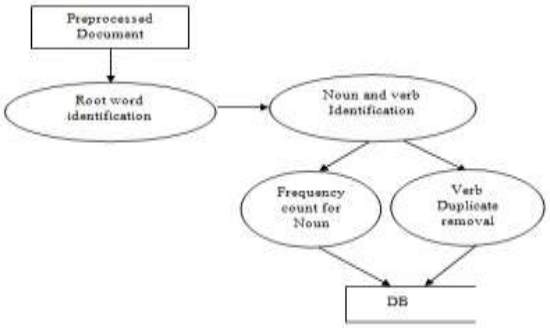

Figure 7 Extraction of Text from Documents

V. Conclusion

The main objective of this thesis is to design and implement semantic oriented methodology to improve recall and precision in IR. Towards the end semantics has been utilized at various linguistic levels and at various stages of IR process. At the word level, additional semantics has been used to improve the word sense disambiguation algorithm, resulting in more accurate sense determination of the indexed words. Incorporation of semantically related sentence level constituents as part of the index structure allowed enriched representation of the document. A concept oriented logical level representation of the document also served as enrichment to the document representation, in addition to allowing inference about the contents of the document. In this way, semantics has been introduced at word level, sentence level and document content level to enhance the representation of the document resulting in improved IR performance. Therefore, a concept based WSD algorithm has been devised in this thesis to improve the performance of Information Retrieval.

In this thesis, we also proposed a model based on an ontology-based scheme for the automatic annotation of Tamil documents and a retrieval system. The retrieval model is based on an adaptation of the classic vector-space model, including an annotation weighting algorithm, and a ranking algorithm. The approach introduced is fairly simple yet highly effective in annotating documents automatically.

Here, Semantic search is combined with conventional keyword-based retrieval to achieve tolerance to knowledge base incompleteness. Experiments are shown where this approach shows clear improvements with respect to keyword-based search. This approach achieved better precision and better recall in this model by introducing ontologies, structured queries, query weightage etc. Structured queries allow expressing more precise information needs, leading to more accurate answers which is possible with a semantic query.

References

[1]. Pablo Castells, Miriam Ferna´ndez, and David Vallet, An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval, "IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 2, February 2007".

[2]. Tao Jiang,Ah-Hwee Tan, Senior Member, IEEE, and KeWang, Mining Generalized Association of Semantic Relation from Textual Web Content. "IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 2, pp.32-35, February 2007".

[3]. V. Christophides, G. Karvounarakis, D. Plexousakis, and S.Tourtounis, "Optimizing Taxonomic Semantic Web Queries Using Labeling Schemes," J. Web Semantics, vol. 1, no. 2, pp. 207-228,2003.

[4]. M. Cristani and R. Cuel, "A Survey on Ontology Creation Methodologies," Int'l J. Semantic Web and Information Systems, vol. 1, no. 2, pp. 49-69, 2005.

[5]. W.B. Croft, "Combining Approaches to Information Retrieval," Advances in Information Retrieval, pp. 1-36, Kluwer Academic, 2000.

[6]. S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R.Harshman, "Indexing by Latent Semantic Analysis," J. Am. Soc.Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[7]. S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T.Kanungo, K.S. McCurley, S. Rajagopalan, A. Tomkins, J.A.Tomlin, and J.Y. Zien, "A Case for Automated Large Scale Semantic Annotation," J. Web Semantics, vol.1, no.1, pp.115-132, 2003.

[8]. S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing," Web Intelligence and Agent Systems, vol. 1, nos. 3-4, pp. 219-234, 2003.

[9]. A. Go´mez-Pe´rez, M. Ferna´ndez-Lo´pez, and O. Corcho, Ontological Engineering. Springer-Verlag, 2003.

[10]. N. Guarino, C.Masolo, and G.Vetere, "OntoSeek: Content-Based Access to the Web," IEEE Intelligent Systems, vol. 14, no. 3, pp. 70- 80, 1990.

[11]. P. Castells, M. Ferna´ndez, D. Vallet, P. Mylonas, and Y. Avrithis, "Self-Tuning Personalized Information Retrieval in an Ontology- Based Framework," Proc. First Int'l Workshop Web Semantics (SWWS '05), 2005.

[12]. Alessi, R.S., Vang L , and Voorhees. W B (1993). Farmbook: A whole-farm information system.

[13]. Amit Bagga, Joyce yue chai and Alan W.Biermann. (1997) The Role of WordNet in the Creation of a Trainable Messege Undestanding System, Proceeding on Artificial Intelligence and the Eighth Innovative applications of Artificial Conference

[14]. Anick P G and Flynn R A (1993), "*Integrating a dynamic lexicon with a dynamic full-text retrieval system"*, In Proceedings of the 16th ACM SIGIR, pp.136-145, New York

[15]. Anick, Shivakumar Vaithyanathan, (1994) Exploiting clustering and phrases for context-based information retrieval, Proceedings of the 20th Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval**...** 295-303.

[16]. *AT&T Bell Laboratories, Murray Hill, New Jersey 07974-2070*

[17]. Ballesteros L & Croft W B (1997). Phrasal Translation and Query Expansion Techniques for Cross-Language

[18]. Bartell (1992) Brian T.; Cottrell, Garrison W. and Belew, Richard K. *Latent Semantic Indexing is an optimal special case of multidimensional scaling.* SIGIR Forum (ACM Special Interest Group on Information Retrieval), p. 161-167

[19]. Berenci, E., Carpineto, C., and Giannini, V. (1998). Improving the effectiveness of Web search engines using selectable views of retrieval results. Journal of Universal Computer Science, vol. 4, n. 9, pp. 737-747

[20]. Bordogna, et. al. (1994) *An extended fuzzy linguistic approach to generalize Boolean information retrieval.* Information Sciences, Applications, Vol.2 (3), p. 119-134

[21]. Brill E (1992). A Simple Rule-Based Part Of Speech Tagger. In *Proc. ANLP-92, 3rd Conf. on Applied Natural Language Processing*.

[22]. Bruce and Wiebe, 1994; Rigau et al., 1997) information retrieval (Krovetz and Croft, 1992) information extraction (Cowie et al., 1993) and text coherence (Kozima and Furugori, 1993)