Acquiring Evolving Semantic Relationships for WordNet to Enhance Information Retrieval

Ms.D.Akila[1] Dr. C.Jayakumar[2]

[1]Research Scholar, Bharathiar University, Coimbatore, Tamilnadu
Asst.Professor, Dept. of computer Applications & IT,
Guru Shree ShantiVijai Jain College for Women, Chennai, Tamilnadu
Email:akiindia@yahoo.comcontact-no: 09962278701

[2]Research Supervisor, Bharathiar University, Coimbatore, Tamilnadu
Professor, Department of Computer Science and Engineering,
R.M.K Engineering College, KavaraiPettai- 601206, Tamilnadu
Email: cjk.cse@rmkec.ac.incontact no: 09884217734

Abstract :Lexical Knowledge base such as WordNet has been used as a valuable tool for measuring semantic similarity in various Information Retrieval (IR) applications. It is a domain independent lexical database. Since, the quality of semantic relationship in WordNet has not upgraded appropriately for the current usage in the modern IR. Building the WordNet from scratch is not an easy task for keeping updated with current terminology and concepts. Therefore, this paper undergoes a different perspective that automatically updates an existing lexical ontology uses knowledge resources such as the Wikipedia and the Web search engine. This methodology has established the recently evolving relations and also aligns the existing relations between concepts based on its usage over time. It consists of three main phases such as candidate article generation, lexical relationship extraction and generalization and WordNet alignment. In candidate article generation, disambiguation mapping disambiguates ambiguous links between WordNet concepts and Wikipedia articles and returns a set of word-article pairings. Lexical relationship extraction phase includes two algorithms, Lexical Relationship Retrieval (LRR) algorithm discovers the set of lexical patterns exists between concepts and sequential pattern grouping algorithm generalizes lexical patterns and computes corresponding weights based on its frequencies. Furthermore, Sequential Minimal Optimization (SMO) selects the suitable good pattern using the optimal combination of weight of lexical patterns and page count based concurrence measures. WordNet alignment phase establishes a new relationship that is not available in WordNet and also aligns the existing patterns based on computed weight. Experimental results illustrate that the proposed approach better than existing mechanisms on benchmark datasets and achieves a correlation value of 0.87. Moreover, the extended WordNet returns high accuracy results in query expansion.

Keywords *:* Semantic Similarity , Disambiguation Mapping, Sequential Pattern grouping, Sequential minimal Optimization , lexical pattern.

I. INTRODUCTION

Owing to the exponential growth of the Internet, IR applications become the most popular tool to get relevant results from the World Wide Web (WWW). For ambiguous queries, it does not return relevant results as required by the user. The IR process must meet following two requirements. How to provide the related pages based on user interest and how to rank the potentially related pages according to their relevance. The accuracy of returning results in IR applications, mainly relies on the measurement of semantic similarity. Therefore, Semantic similarity is a fundamental and effectiveness concept to get the most relevant results. It measures the relatedness between two concepts which are not lexicographically similar and also explains the degree to which the words are related together [19]. Semantic similarity measurement plays an effective role in many IR applications [10].

It is important to mention that the majority of the existing web applications is planned to undergo a reengineering process , in order to take advantages like reusability, interoperability, integration etc. [8]

Semantic relationship between words of specific concepts is listed manually in lexical Ontologies such as the popular open-domain vocabulary,WordNet [14].It is quite large and offers wide coverage of lexical relations. It consists of a set of interconnected nodes as concepts and the links connecting the nodes as various types of relations between the concepts, such as synonymy, homonymy, Holonymy, Hypernymy and etc. Various semantic similarity measures such as Information Content [12], the path length, and overlap based measure are available in the literature to measure the semantic similarity [11]. Most of the researches approve WordNet as a suitable source to measure semantic similarity [9] [13] [4].Even though, many potentially useful

relations that arise over time based on the usage are missing in WordNet. However, these new evolving relationships are not necessarily appropriate for a general purpose domain independent lexical database. Thus, the proliferation of Senses and lack of new evolving relations between concepts is considered as the main shortcomings for Natural Language Processing (NLP) and IR applications. "Apple" is an example, and is recurrently connected with "computer" on the WWW. However, this Sense of "apple" is not available in WordNet. Therefore, it returns irrelevant results for the ambiguous query "Apple". Since WordNet is a precompiled lexical database, newly updated relationships are not present in it. It is time consuming and an expensive process to capture a new relationship manually and assign these new Senses to existing words.

The main aim of this work is to provide such links between the concepts in WordNet. Over the period of time relationships represented between the concepts in WordNet may get weakened, or new relationships may have high similarity that is not presented in WordNet. The core work of this paper is to align newly evolving strong relationships and already existing weak relationships over time between concepts in WordNet. This work proposes an enhanced methodology to enrich a WordNet with new relationships semantically using the information extracted from the knowledge web resources such as the Wikipedia and the Web search engine. These two resources are being constantly updated according to the user interest over time and across domains. The search engine provides page count and snippets which are very useful to extend existing WordNet.

The contributions are summarized as follows:

- This work proposes a more effective approach to leverage the new relationship into WordNet using two web resources such as Wikipedia and web search engine.
- Disambiguation Mapping performs mapping between Wikipedia pages and WordNet concepts to disambiguate the user ambiguous query effectively.
- It proposes a shallow lexical relationship extraction algorithm to extract different lexical patterns between two words over returned snippets.
- Sequential Pattern grouping (SPG) Algorithm groups the number of extracted lexical patterns that refers the same semantic relation.
- Sequential minimal Optimization (SMO) classifier integrates different web-based similarity measures to select the suitable pattern among the large number of extracting lexical patterns.
- This approach is shown to be fruitful and also infers a number of useful relationships than existing approaches.

A. Paper Organization

This paper is organized as follows. Section 2 provides the discussion of existing works. Section 3 portrays the proposed approach for acquiring new relationship in WordNet from web resources. Section 4 provides the experimental evaluation of the proposed approach and the conclusions are drawn in Section 5.

II. RELATED WORK

There has been a large number of works available in the existing literature to extend WordNet using web resources either Wikipedia or web search engine. The main contribution of this work is to utilize both resources to infer a new relationship between concepts in WordNet arises over time. Agirre et al. (2000) proposed an automatic method to enrich the concepts in WordNet by exploiting the information in WWW [1]. For each Sense of the concept in WordNet, it constructs topic signature vector, i.e., a set of keywords that are topically related to the concept. To deal the increasing Sense proliferation, this approach builds a hierarchical cluster of concepts that lexicalize Senses of a given word. Ruiz-Casado et al. (2005), proposed a method to enrich the existing lexical semantic network automatically using Wikipedia entries. It performs the mapping to associate each Wikipedia category automatically with concepts in a WordNet to identify new concepts from Wikipedia. It utilizes a text overlap based similarity measure to compute the similarity between page's text and Synsets using the vector space model. This approach computes the similarity only considering the text overlap based measure, without utilizing the link structure of Wikipedia. Another drawback is that, it mainly concerns to enrich WordNet with Wikipedia entries which are not found in WordNet but does not concentrate to evolve a new relationship between already existing concepts in WordNet [22].

Latterly, Ruiz-Casado et al (2006) addressed the problem of automatically identifying the semantic relationship from the text information. It presents an automatic procedure to extend the semantic network with the new relationship extracted from Wikipedia. This paper presented a new edit distance algorithm that automatically generalizes the identified lexical patterns without human intervention. This approach extracts four types of relationships such as Hyperonymy, Hyponymy, Meronymy and Holonymy [20]. Suchanek et al., (2007) presented extensible ontology called as YAGO that links to sources such as WordNet and Wikipedia with near perfect accuracy. It is a combination of both rule-based and heuristic method to extract the contents from

Wikipedia and unified with WordNet. This ontology covers a wide range of relations heuristically with high accuracy. However, this approach does not deal with the ambiguity while linking both resources, and only solves ambiguous mappings manually [21].

Ponzetto and Navigli (2010) present a knowledge rich methodology to maximize the text overlap between WordNet taxonomy and category graph extracted from Wikipedia. Due to the wide coverage of Wikipedia, it is associated with WordNet Senses and topical relations. The newly identified semantic associative relations connecting Wikipedia pages are transferred to the WordNet Senses to generate a new extended WordNet++ suitable for the word Sense disambiguation. However, this method performs mapping by considering only the titles and categories rather than the content of the Wikipedia pages [18]. Niemann and Gurevych (2011) proposed a novel two-step approach that automatically aligns WordNetSynsets and Wikipedia articles to extend WordNet with high coverage and quality. This approach increases the efficiency of WordNet and Wikipedia through effective alignment using a word overlap measure. But this approach does not address the ambiguity problem arises according to time [15]. Fernando and Stevenson (2012) perform the mapping between WordNetSynset and Wikipedia articles to add a new relation between concepts in WordNet [7]. All of the existing approaches enrich the existing concepts in WordNet with new concepts and named entities using web resources. However, none of the approaches tried to create the new links between topically related concepts which arises over time. Further, they did not consider the procedure to disambiguate the ambiguous links in WordNet. Hence, the proposed work considers these shortcomings in existing approaches as the main criteria while designing the suitable framework to enrich WordNet for IR and the word Sense Disambiguation Task.

III. IDENTIFICATION OF NEWLY EVOLVING RELATIONS IN WORDNET

In order to establish new relations and align existing relations in WordNet based on the usage, this proposed approach rely on three phases such as

i) *The Candidate Generation*: Disambiguation mapping is performed to acquire the new Senses of the concept in WordNet from Wikipedia pages.

ii) *Lexical pattern Extraction*: Extraction of lexical patterns and the computation of its corresponding weight by exploiting two kinds of useful information such as page counts and snippet returned from the web search engine.

iii) *WordNet Extension & Alignment:*Create newly evolving relations between concepts and also aligns relations based on its computed weight. Thus, Extended WordNet is aligned according to the strength of the relationship based on the current usage in the modern world. Figure1 shows the overview of the proposed approach.

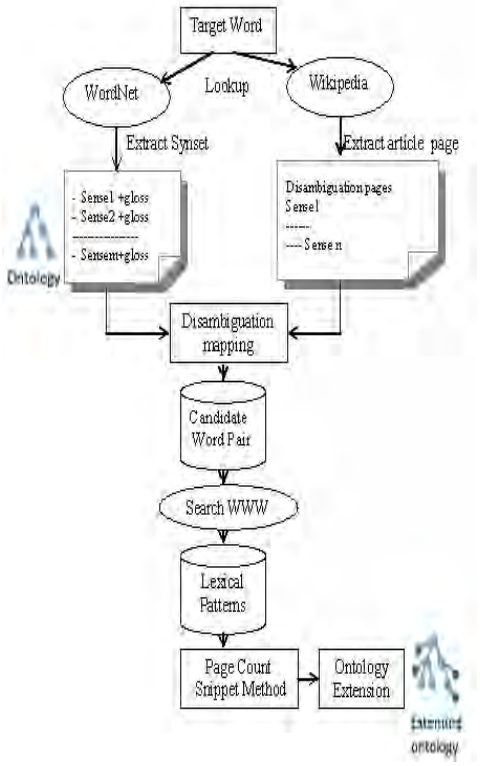

Figure 1: An extension of the newly evolving relationship in the WordNet

A. Automatic Acquisition of candidate articles from Wikipedia

Wikipedia is a collaborative, highly structured web encyclopedia composed of a number of pages. It provides the knowledge about a specific concept or named entities and also explicitly represents the different Senses of the concept according to the context. In this work, we download a Wikipedia dump from the Wikipedia download site available in online. And then, JWPL (Java WikiPedia Library)**,** a Java-based application programming interface used to extract the required pages in Wikipedia dumb. For a target word, Wikipedia consists of three different kinds of pages such as article pages, Link or redirect pages and Disambiguation pages [17].

- Article pages are the pages whose title matches the target word. These pages are available at pages-articles.xml.bz in Wikipedia dumb.
- Link or redirect pages direct the query to the concrete page that contains the actual information about the query. For example the word 'apple' redirects to 'fruit'. These pages are availableat pagelinks.sql.gz in Wikipedia dumb.
- Disambiguation pages contain a link to the number of possible concepts that the ambiguous entity could be pointed to. They are specially designed to disambiguate a number of similar words that refer to a single ambiguous term. The unique identifier of a disambiguation page consists of the name of the ambiguous entity attached followed by the tag disambiguation. These pages are available at Disambiguation_pages ~/enwiki/20081013 in Wikipedia dumb.

1)Disambiguation Mapping

Disambiguation mapping associates each WordNet entry into its corresponding Wikipedia entry, so that the newly evolved relationship in Wikipedia can be easily determined. The work in [20] also took an effort to reconstruct the WordNet with newly evolving relations, but it takes into account of all Wikipedia pages (article, link, Redirect, Disambiguation pages) related to the target word in an entire Wikipedia Corpus. Therefore, it becomes a tedious process to disambiguate the ambiguous links in WordNet and also difficult to identify the different senses of the word in a large dimensional space. This work considers only the disambiguation pages to reduce the large dimensional space. In a target ambiguous word, we retrieve the Wikipedia disambiguation pages whose title is in the form of the ambiguous word with parenthetical explanation disambiguation. These pages are more important for this work because it is very helpful to identify the new relations that are not available in WordNet.In WordNet, the ambiguous word "apple" associated with two different lexical units (Senses).In this work, we extract Synsets using Java API for WordNet searching (jaws) that provides the ability to retrieve data from the WordNet database. The Synset for an apple is extracted as follows

The noun apple has 2 senses (first 1 from tagged texts)

1. (2) apple -- (fruit with red or yellow or green skin and sweet to tart crisp whitish flesh)
2. apple, orchard apple tree, Malus pumila -- (native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits)

Figure 2 : Two Senses of word "apple"

On the other hand, the disambiguation page in Fig. 3 lists more than 5 Senses for Apple: Companies, plant and plant parts, Television, music and Technology. Each of these Senses summarizes the list of all distinguished meanings of an ambiguous word along with short descriptions and also consists of a link to the corresponding Wikipedia article. Additionally, the disambiguation page also listsdifferent Senses such as Apple Corps (multimedia corporation), Apple Inc., (consumer electronics and software company) Apple Bank (an American bank in the New York City area) etc.,. Since these Senses are not available in WordNet, so these mappings identify the additional Senses of the word rather than ordinary Sense in WordNet.

Figure 3: A disambiguation page of the ambiguous word "apple"

2) Mapping Algorithm

Ponzetto andNavigli et al., presented a disambiguation algorithm [18] that performs a mapping to identify the disambiguation contexts between two resources. This algorithm maps each candidate article in the Wikipedia page with its matching Synset in WordNet. On the contrary, this work proposes a novel mapping algorithm to link WordNet Senses (S_{wnet})and senses in Wikipedia disambiguation pages (S_{wiki}). It is mainly used to determine the number of additional senses available in Wikipedia pages rather than in WordNet. The following procedure is performed:

Step1: At the initial stage, the mapping $\Omega(w_p, w_n)$ is empty, therefore the link between Wikipedia Sense (S_{wiki}) and WordNet Sense (S_{wnet}) isϕ;

Step 2 : If the number of Wikipedia Sense and WordNet Sense is equal to one, then there is no disambiguation page, i.e. the word is monoecious. Therefore, the Wikipedia article is used as a candidate for mapping.

Step 3: If the number of Wikipedia Senses is greater than the number of WordNet Senses then the external candidate articles on the Wikipedia page is added as a new Sense over the ambiguous word in WordNet. The conditional joint probability $p(S_{wiki}|w_p)$ is computed to select the suitable Wikipedia Sense (S_{wiki}) from the Wikipedia disambiguation page wp. The Sense with the highest score is computed based on the intersectionof the disambiguation context of Wikipedia Sense in its disambiguation page. The result of the mapping from WordNet to Wikipedia isa set of word-article pairings. The pseudo code of the mapping algorithm is as follows

Input: Wikipedia Sense S_{wiki}, WordNet Sense S_{wnet}, Wikipedia Page w_p,
Wordnetw_n
Output: (W_{ambi}, $W_{candidatel}$). (W_{ambi} ,$W_{candidate2}$)……..(W_{ambi}, $W_{candichtev}$)

```
for each w_p ϵ S_wiki
        Ω(w_p,w_n) = ϕ // initialize mapping
for each w_p ϵ S_wiki
if ( |S_wiki| = |S_wnet| =1) then
        Ω(w_p,w_n) = w_n ; // monosemous
else if (( |S_wiki| > |S_wnet| ) then
    Ω(w_p,w_n) = max p(S_wiki,w_p) //polysemous
    p(S_wiki | w_p) = argmax score (S_wiki, w_p)
    score (S_wiki,w_p) = | di sambi (S_wiki) dis∩ ambi(w_p) | ;
    return Ω (w_p,w_n,) = (w_ambi, w_candidate); where v =1,2,3....
        end ;
```

Algorithm 1: Disambiguation Mapping

B. Searching at WWW

Disambiguation mapping returns a word-article pairings containing a seed list of word pairs such as {(apple, fruit), (apple, tree), (apple, consumer electronics), apple, multinational company)} and so on. Then, every word pair is submitted as an input to the web search engine. This work uses the Google Search API that provides the search results from the generic search engines. The word pairs are sent to the search API which fetches two kinds of useful information such as the page count and a snippet for the given query over the web. Page counts for the query w_{ambi} AND $w_{candidate}$ is defined as the approximation of co-occurrence of two words on the web. This work uses snippets returned by the search engine for the combination of query with both "apple and computer". It presents two algorithms such as the LRR Algorithm to extract lexical patterns from collected snippets and SPG Algorithm computes the frequency of the occurrence of extracting lexical patterns between the word pair.These extracted lexical patterns are integrated as a newly evolved semantic relationship for the ambiguous word in WordNet automatically.

1) Page count co-occurrence measure

Bollegala et al., used four co-occurrence based measure to compute the semantic similarity in their work in [2]. Our approach additionally considers the Normalized Google distance, therefore, totally five web measures such as Web-Jaccard, Web-Overlap (Simpson), Web-Dice, Web-PMI (Point-wise mutual information) and Normalized Google Distance (NGD) are used to calculate robust semantic similarity by means of page counts. Page counts for the ambiguous word (w_{ambi}), candidate word ($w_{candidate}$) are represented as PC (w_{ambi}) and PC($w_{candidate}$) respectively. The WebJaccardcoefficient between word pairs (w_{ambi}, $w_{candidate}$) using page count is represented as follows

$$WebJaccard(w_{ambi}, w_{candidate}) = \begin{cases} 0 & if\left(PC(w_{ambi} \cap w_{candidate})\right) \leq \gamma \\ otherwise & \\ \dfrac{PC(w_{ambi} \cap w_{candidate})}{PC(w_{ambi}) + PC(w_{candidate}) - PC(w_{ambi} \cap w_{candidate})} & \end{cases} \quad ---------(1)$$

The web Overlap co-efficient between a word pair (w_{ambi},$w_{candidate}$) using page count is

$$WebJaccard(w_{ambi}, w_{candidate}) = \begin{cases} 0 & if\left(PC(w_{ambi} \cap w_{candidate})\right) \leq \gamma, \\ otherwise & \\ \dfrac{PC(w_{ambi} \cap w_{candidate})}{\min\left(PC(w_{ambi}), PC(w_{candidate})\right.} & \end{cases} \quad ------(2)$$

The Web Dice co-efficient between a word pair (w_{ambi},$w_{candidate}$) using page count is

$$WebDice(w_{ambi}, w_{candidate}) = \begin{cases} 0 & if\left(PC(w_{abmi} \cap w_{candidate})\right) \leq \gamma, \\ otherwise & \\ \dfrac{2PC(w_{ambi} \cap w_{candidate})}{PC(w_{ambi}) + PC(w_{candidate})} & \end{cases} \quad --------(3)$$

The web pointwise mutual information measure between a word pair [18] using page count is

$$WebPMI(w_{ambi}, w_{candidate}) = \begin{cases} 0 & if\left(PC(w_{ambi} \cap w_{candidate})\right) \leq \gamma, \\ otherwise & \\ \dfrac{Log_2\, PC(w_{ambi} \cap w_{candidate})/N}{PC(w_{ambi})/N + PC(w_{candidate})/N} & \end{cases} \quad ----------(4)$$

The Normalized Google Distance measure [19] between a word pair using page count is

$$NGD(w_{ambi}, w_{candidate}) = \frac{\max\{\log PC(w_{ambi}), \log PC(w_{candidate})\} - \log PC(w_{ambi}, w_{candidate})}{Log\ N - \min\{\log PC(w_{ambi}), \log PC(w_{candidate})\}} \quad ---(5)$$

In the equation (1) (2) (3) (4) and (5), PC($w_{ambi} \cap w_{candidate}$) denotes the page count of conjunction Query of the ambiguous word in WordNet, and its corresponding candidate article in Wikipedia. N is the number of indexed documents by a web search engine. There is a possibility for the occurrence of two words due to high noise and a large availability of web data in some pages, even though there is no relationship between them. Therefore, if the page count of a query is less than γ, the value of web co-occurrence coefficient value will be set to zero.

2) Lexical Relationship Retrieval (LRR) Algorithm

Web search engine returns a text snippet for the conjunctive query given as a form of w_{ambi} ***** $w_{candidate.}$ The snippet is a short description consists of less than 20 words, including queried word pairs, and its corresponding semantic relation between them. It provides more precise information related to its local context. In this approach, we collect the set of lexical patterns exists between each word pair. Furthermore, the

representation of the relationship between them as a triple (w_{ambi}, $w_{candidate}$, R) consists of two related concepts and its corresponding relationship itself. This paper proposes a shallow lexical relationship extraction algorithm by extending the Pattern extraction algorithm in [20] to extract different lexical patterns. Since, the semantic relation between two pairs can be represented using different lexical patterns. This algorithm acquires a set of lexical relationship exists between a word pair, and also computes the frequency of occurrence of patterns in returned snippets. It is a light-weight algorithm that does not require any complex interpretation modules or inferred procedure to extract lexical patterns. It extracts the subsequence that consists of a word pair, and its corresponding relationship between them, from the snippet that meets the following constraints.

- A subsequence should have an occurrence of the word pair exactly one time.
- The length of a Subsequence does not exceed L_{max}words.
- The stemming operation allows to skip only less than c number of words consecutively.

Input : Snippet S returned for the query w_{ambi} ** $w_{candidate}$**
Output: Set of lexical patterns
for each snippet S do
Read Snippet (S)
If (W_{ambi}ϵ S &&$W_{candidate}$ ϵ S)
Perform Stemming (S);
Endif
L (S) = Number of words (S)
compute n(w_{ambi}) = Number of occurrences (S, w_{ambi});
compute n($w_{candidate}$) = Number of occurrences (S, $w_{candidate}$);
if (L(S)>L_{max}) then
if (n(w_{ambi}) >1 &&n($W_{candidate}$) >1)
allow (w_{ambi}, $W_{candidate}$) only one time
returnsubseq;
end if;
endif;
else if (L(S)<L_{max})
Discard S and read next snippet S;
end if;
for each subseq do
if (w_{ambi} ϵ subseq&&$w_{candidate}$, ϵ subseq)
return Set of patterns from subseq;
end for;
end for;

Algorithm 2: Lexical Relationship Retrieval (lRR) algorithm

3)Sequential Pattern grouping (SPG) Algorithm

As demonstrated before, the proposed LRR algorithm automatically extracts a large number of lexical patterns exist between the word pair. LRR algorithm yields numerous unique patterns. Hence, most of the patterns occur less than 10 times. Existing pattern clustering algorithm undergoes a pair-wise comparison to cluster the set of the same patterns. However, it is a time consuming process due to the pair-wise comparison between the large number of unique patterns. Therefore, this work undergoes a pattern grouping in a sequential nature that avoids the time complexity in pair-wise comparisons. It effectively groups the extracted different lexical patterns that correspond to the same semantic relation. The extracted i number of lexical patterns that co-occur between the word pair (w_{ambi}, $w_{candidate}$) are grouped into $G_j$cluster. The frequency $\omega(lp_i)$ of lexical pattern lp_i between a word pair (w_{ambi}, $w_{candidate}$) is the total occurrence of the lexical pattern lp_i computed as follows

$$\omega(lp_i) = \sum f\,(w_{ambi}, w_{candidate}, lp_i) - - - - - - - (6)$$

Input LP = (1p1, 1p2.....1p$_m$),n) & threshold λ

Output Set of Pattern Group G= {g*gg***.....g^k};**

For each lexical pattern & word pair (w$_{ambi}$, w$_{candidate}$)

Compute frequency ω (lp$_i$)=Σ*f*(w$_{ambi}$, w$_{candidate}$,Ip$_i$)

SORT (LP) according to w (LP)

Initialize G =ϕ;

Compute similarity = cosine (lp$_i$,lp$_j$)

if (similarity > λ) then

Insert lp$_i$ →group g* has lp$_j$;

g* becomes [lp$_i$,lp$_j$]

else

Insert lp$_i$→new duster g;**

end if;

Algorithm 3: Sequential Pattern Grouping (SPG) Algorithm

The description of the algorithm is explained as follows

Step 1: The SORT method sorts the set of lexical patterns {lp$_{1,}$ lp$_{1.....}$ lp$_m$} into descending order according to the frequency of its occurrence in all word pairs. Therefore, the high frequency patterns (popular) occur at the beginning, whereas low frequency gets transferred to last.

Step 2: Initialize the Group G as an empty set.

Step 3: Compute the cosine similarity between the lexical pattern lp$_i$ and lp$_{j.}$ If it has high similarity greater than λ, then insert lpi into the group g* that has lp$_{j.}$ Otherwise, insert lpi into a new group g**.

Step 4 : Follow the step 3 for all the extracted lexical patterns.

4) Measuring Semantic Similarity using SMO-SVM

Bollegala et al., follows a machine learning based approach using SVM to compute the semantic similarity score by integrating both page count based co-occurrence measure and a group of lexical patterns extracted from snippets [2]. However, it is not an easy task to train an SVM classifier because of large quadratic programming problem occurs in the case of numerous unique patterns. It also takes a more time while trained with a large training set that affects the processing speed. Therefore, this work uses sequential minimal Optimization (SMO) [16] to train an SVM that breaks the large quadratic problem into a number of small possible problems. SMO-SVM can handle large training sets because the memory requirement is linear to the size of the training set. It achieves greater training speed even for a large training set by completely eliminating the time and computational complexity. In this approach, SMO-SVM is trained using two types of training sets such as positive training set with synonym word pairs and negative training set with non-synonym word pairs to determine the lexical pattern that has a high semantic similarity between the word pair. After training SVM with training sets, feature vector f* are generated for both page count based co-occurrence measure and lexical patterns of snippet extraction. Therefore, the feature vector f*(w$_{ambi}$,w$_{candidate}$) of all page count based co-occurrence measure (Jaccard, Dice, Overlap, PMI and NGD) for a pair of words(w$_{ambi}$,w$_{candidate}$) are represented as (K+1)st, (K+2)nd, (K+3)rd, and (K+4)th and (K+5)th dimensional vector respectively. The feature vector for lexical patterns is the weighted sum of all patterns in the cluster G$_j$ is defined as follows

$$f(w_{ambi}, w_{candidate}) = \sum \varphi(lp_i, G_j) f(w_{ambi}, w_{candidate}) where\ lp_i \epsilon\ G_j \quad -----(7)$$

Where ω$_{pic}$ is the weight of the pattern up in the cluster q is represented as follows

$$\varphi(lp_i, G_j) = \frac{\omega(lp_i)}{\sum \omega(h)}\ where\ h \in G_j \quad -----(8)$$

Therefore, the semantic similarity between the word pair (w$_{ambi}$,w$_{candiate}$) is computed using posterior probability as in (9) when feature vector f* belongs to the synonymous training set

Semantic Similarity(wambi,wcandiate) = P (f* | synonymous set) ------ (9)

Using (9), compute the semantic similarity for all the word pair{(w$_{ambi}$, w $_{candidate1}$), (w$_{ambi}$, w $_{candidate2}$) ……(w$_{ambi}$, w $_{candidaten}$)}. Then, realign the semantic relation between concepts in WordNet using the semantic similarity computed from the page count snippet method.

C. Ontology Extension using New Relationship

Ontology Extension is an automatic process includes three main tasks such as integration of new concepts into WordNet, instantiate a newly evolved relationship between existing concepts, and aligns the existing relationship between concepts. This work intends to extend the lexical ontology WordNet by performing the last two tasks. Generally, WordNet is represented as $w_n(ogl) = (N,R)$ where N is the set of nodes represent the word or concepts, R is the existing lexicon semantic relation between them. The ontology for the word Apple in WordNet is in the form of a figure 4. The extended ontology represented as w_n (Xtn) = N, R, R_n that encompasses new context relations derived from the web search engine. The semantic similarity is computed for all the lexical patterns exist between word-article pairings (W_{ambi}, $W_{candidate}$) extracted from both knowledge resources such as Wikipedia and WordNet. Then, the lexical pattern with a higher semantic similarity score is emerging as a new relation between two concepts (w_{ambi}&$W_{candidate}$).

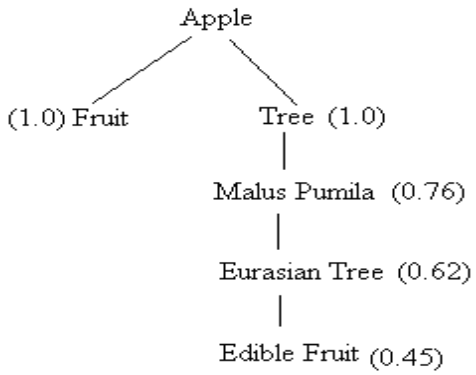

Figure 4: Apple concept in WordNet

In an ambiguous word, Wikipedia has a number of Senses that are not available in the WordNet. This causes the lesser relationship and Senses representation of the word in WordNet. For overcoming this, this approach links the new Senses to the concepts and its relationships to WordNet taxonomy through our novel method. For example, the ambiguous word “apple” (w_{ambi}) has only two Senses in WordNet such as fruit and plant, whereas the “apple” (w_{ambi}) has more than seven Senses in Wikipedia such as consumer electronics, company, film, an American bank and so on. This work intends to explore these new Senses of the ambiguous word into WordNet for achieving effective information retrieval. Therefore, the semantic similarity between the apple (w_{ambi}) in WordNet and the company ($w_{candidate}$) in Wikipedia are computed using our approach, then the lexical pattern with high semantic similarity score are extended as a new relation between the two concepts “apple” and “company” in WordNet. Therefore, lexical patterns with high score are applied to discover new relationships between the concepts which are already belong to the ontology. On the other hand, the importance of the relationship between two concepts diminishes over time, i.e., less semantic score between those concepts. Therefore, this work aligns the relationship between them according to the computed semantic similarity using the web. For instance, compute the number of occurrences of relations for the already existing relations between “apple” and “fruit” in WordNet. According to the frequency, weight is assigned to the relation between it. Thus, the existing relationship between concepts is aligned according to the semantic similarity computed from the web.

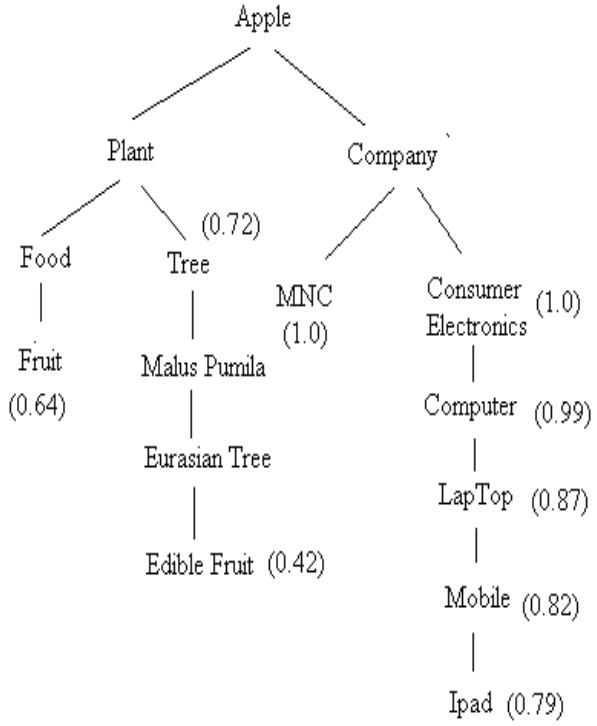

Figure 5: Apple concept in extended WordNet.

IV. EXPERIMENTAL EVALUATION

Three semantic similarity evaluation approaches are available in the literature to estimate the performance of the proposed work. First, theoretical examination of the proposed measure is carried out to evaluate its effectiveness based on mathematical properties. Second, the proposed semantic measure is compared with humanratings in three well-known benchmark data sets such as Miller-Charles (MC), Word Similarity-353 (WS) and Rubenstein-Goodenough (RG). Third, the proposed measure is applicable to real-world applications to compute the similarity and then evaluate the measures with respect to their performance in applications such as NLP, WSD, and IR systems. This work utilizes a second and third set of experiments to evaluate the performance of the proposed semantic measures.

A. Evaluation using Benchmark Datasets

Initially, we evaluate the similarity measure of the proposed approach against the benchmark dataset called as Miller-Charles (MC) that consists of similarity judged by human experts. Pearson Product-moment Correlation Coefficient is used to determine the consistency between the proposed semantic similarity measures and human ratings in a benchmark data set.

B. Benchmark Dataset

To measure semantic similarity, two-class SMO-SVM is trained with synonym word pairs and non-synonyms word pairs from the extended WordNet. In order to evaluate the proposed semantic similarity score, the reliable benchmark data set called as standard Miller-Charles dataset, which is having 28 word-pairs and human rated semantic similarity are taken as test data. The similarity between word pairs is rated on a scale from0 (no similarity) to 1 (perfect synonymy).

1) Experimental Results

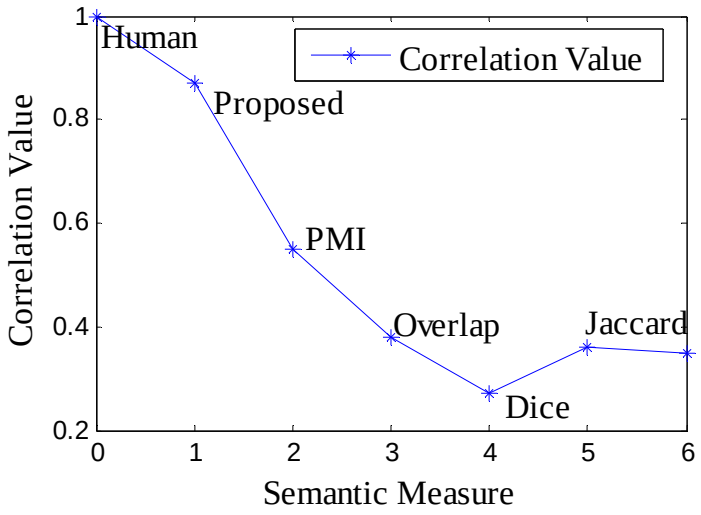

Figure 6: The Correlation of proposed semantic similarity measure

Figure 6 shows the correlation coefficient of a proposed approach with other existing web based measures such as web Jaccard, web Dice, web Overlap and web PMI. It illustrates that the proposed semantic measures achieve a correlation value of 0.87 as that is nearer to the human assessment. Therefore, we can believe that our proposed measure is better than other existing web based semantic similarity measures.

C. Automatic Query Suggestions (AQS) in IR

Query suggestion is an ultimate feature of the web search engine that provides alternative queries to the user. The search engine returns relevant documents that have more query terms as in the user submitted query (Q). This work proposes a most promising Query

Query suggestion is an ultimate feature of the web search engine that provides alternative queries to the user. The search engine returns relevant documents that have more query terms as in the user submitted query (Q). This work proposes a most promising Query suggestion scheme that improves the usability of web search engines by suggesting a set of related queries to describe the user information more clearly and also to obtain more relevant results even for rare queries in a web search. In our proposed approach, a user submitted query (Q) is reformulated using related terms ($T_{related}$) from the concepts of enhanced WordNet ontology that suggest the potentially related queries even for a new occurring and rarely submitted query.

Conceptual Query Suggestions

Enhanced WordNetused as a knowledge base to perform a clever search with semantic reasoning for high relevant results. It associates a large number of proliferations of Senses acquired from web search engine and semantic relationships are also aligned according the current trend of the people. Therefore, it can handle the ambiguous and exploratory queries by semantically inferring the correct sense of the original query as expected by the user. It returns more accurate results with high precision and recall. Enhanced WordNet is used in existing search engines to reformulate the query and make a intelligent search. The given technique is very efficient and scalable; it is particularly effective in generating suggestions for rare queries and newly occurring queries. It consists of three steps to achieve query expansion:

a)Candidate extraction:

First, extract the number of candidate terms ($T_{candidate}$) for a query based on their semantic relations in an enhanced WordNet such as Synonyms, Hypernyms/Hyponyms, Holonymy etc.

b) Weighting scheme:

To select the related terms ($T_{related}$) for a query, this paper proposes a weighting scheme that selects the related terms based on the similarity between the terms. And then, assign the appropriate weight for the extracted candidate terms, according to the semantic similarity between the candidate terms ($T_{candidate}$) and query terms (T_{query}).

c) SVM classifier:

Last, the candidate terms along with its similarity value are sent as an input to the SVM classifier that classifies the terms into three classes such as high weight (more similar), lightly weight (similar) and less weight (not similar). Furthermore, the term with heavy weight is mostly preferable as a suitable term is appended to the query to form an expanded query (Q_{expand}). It avoids the wrong expansion terms that is not

similar to the original query terms. For example, the user submits as a query as an "Apple price in India" to find the apple Product price. Query Expansion based on wordnet suggest the queries such as "Malwa apple tree", "Apple fruit price in Kashmir" etc., these expanded queries are totally irrelevant for the user expectation. Since, apple has only two senses, such as fruit and plant in the conventional wordnet. On the other hand, Query expansion using enhanced wordnet suggests more relevant queries such as Apple phone price in India, Apple Mac Laptop price, etc. Because, extended WordNet consists of totally 7 senses that are newly added to the wordnet based on the current interest of the user. Therefore, it returns most relevant results as expected by the user.

1) Dataset

This approach uses documents and topics in a large test collection called as TREC-6 to assess the relevance of returned results for regarded query. It consists of totally 556,077 documents with nearly 450 topics. Among that, short queries are formulated using only the title of 150 topics with identifiers 301-450. It adopts a vector space model based text retrieval system, SMART version 11.0 that follows TF-IDF weighting method for information retrieval. It would lead to the best improvement in IR evaluation metrics.

2) Performance assessment of information retrieval

The main purpose of this experiment evaluation is to compare the effectiveness of the proposed query expansion method with one of the existing query expansion methods. The retrieval effectiveness of the IR process is evaluated using the combination of three measures such as user satisfaction and the relevance of retrieved documents in terms of two metrics called as precision and recall.

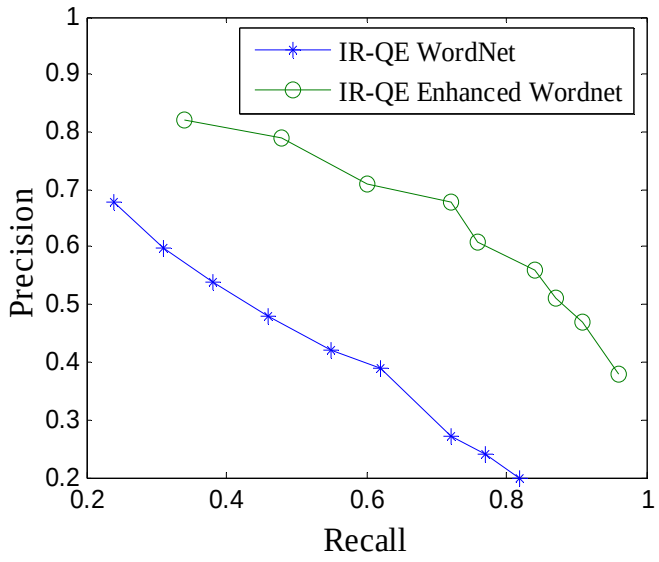

Figure 8: Accuracy of returned results

Figure 8 illustrates that performance of IR with query expansion (QE) based on enhanced WordNet outperforms the IR with query expansion based on WordNet in terms of accuracy in returned results. The proposed query expansion with the appropriate term weighting strategy can exploit the semantic relations in lexical resources to significantly improve retrieval performance. It shows that our scheme has gained better precision than others. In existing systems, the precision of a WordNet is slumped down to the least score, and it can't be retained for the search results. Hence, it reveals that the information retrieval based on enhanced WordNet returns more appropriate results according to the user intent.

Wordnet	Precision	0.68	0.60	0.54	0.48	0.42	0.39	0.27	0.24	0.20
	Recall	0.24	0.31	0.38	0.46	0.55	0.62	0.72	0.77	0.82
Enhanced WordNet	Precision	0.82	0.79	0.71	0.68	0.61	0.56	0.51	0.47	0.38
	Recall	0.34	0.48	0.60	0.72	0.76	0.84	0.87	0.91	0.96

Table 1: Precision and recall of ir-qe based on WordNet and enhanced wordnet.

3)Effect of the Number of Senses

Figure 9 shows that the average precision of returned results via a number of the Senses in the WordNet. From this figure, it is obviously visualized that the retrieval performance of IR-QE based on enhanced WordNet reaches a maximum precision of 0.82 because it effectively integrates the newly evolved senses in the current world. Therefore, it can return more relevant results as expected by the user. Whereas IR-QE based on WordNet reaches only the precision value of 0.67 because the senses available in WordNet is not precise and not upgraded according to the current trend of the people. Therefore, it suggests wrong expansion terms and also returns less relevant results with less accuracy.

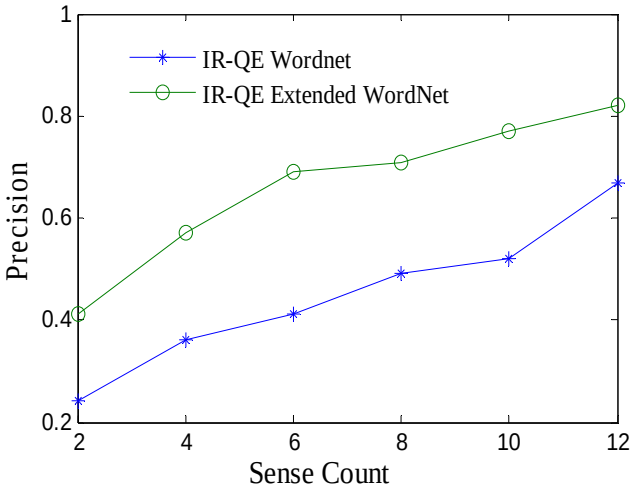

Figure 9: Precision versus the number of senses in WordNet

V. CONCLUSION

This paper proposed an approach that acquired evolving new relationships and also aligned the existing relations in WordNet using web resources such as WordNet and Wikipedia. It presented a lexical relationship retrieval algorithm and sequential pattern grouping algorithm to extract and group the number of lexical relationships from the text snippets retrieved from a search engine. A two-class SMO-SVM was trained to select the suitable pattern that was extended as a new relationship to the ambiguous word in WordNet. Experimental results showed that the proposed method outperforms existing semantic similarity measure and achieves a correlation value close to human ratings in benchmark dataset. Moreover, the proposed method is implemented in real world applications such as IR and QE and also demonstrated that the extended WordNet improved the accuracy of returning results.

REFERENCES

[1] Agirre E., Ansa O., Martinez D., Hovy E., “Enriching WordNet concepts with topic signatures”, Proceedings of the NAACL workshop on Wordnet and other lexical resources:Applications, Extensions and Customizations, 2001.

[2] BollegalaD.,Matsuo Y. and Ishizuka M., “A Web Search Engine-Based Approach to Measure Semantic Similarity between Words”, IEEE Transactions on Knowledge and Data Engineering, Vol.23, No.7, pp. 977-990, 2011.

[3] Bollegala D., Matsuo Y., Ishizuka M., ”Measuring Semantic Similarity between Words Using Web Search Engines”, Proceedings of the 16th International conference on World Wide Web, pp. 757-766, 2007.

[4] Budanitsky A., Hirst G., “Evaluating WordNet-based measures of lexical semantic relatedness”, Computer Linguistics, Vol 32, No 1, pp 13–47, 2006.

[5] Church K., Hanks P., “Word association norms, mutual information and lexicography”, Computational Linguistics, Vol 16, No 1, pp. 22-29, 1990.

[6] Cilibrasi R. and Vitanyi P., “The Google Similarity Distance”, IEEE Transactions on Knowledge and Data Engineering, Vol.19, No.3, pp.370-383, 2007.

[7] Fernando S., Stevenson M., “Mapping WordNetSynsets to Wikipedia articles”, In proceedings of the Eight International Conference on Language Resources and Evaluation, pp. 590-596, 2012.

[8] Fethallah H. and Amine M. ,”Automated Retrieval of Semantic Web Services : A Matching based on conceptual Indexation”,The International Arab Journal of Information Technology, Volume 10 No 1, pp. 61 – 66 ,2013.

[9] Hliaoutakis1 A., Varelas G., Voutsakis E., Petraki E., and Milios E., “Information Retrieval by Semantic Similarity”, International Journal on Semantic Web and Information Systems, Vol 2, No 3, 2006.

[10] Inkpen D., “Semantic Similarity Knowledge and its Applications”, StudiaUniversitatis Babes-BolyaiInformatica, Vol XLV, No 1, pp 11-22, 2007.

[11] Jiang J.,ConrathD.,“Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy” Proceedings of International Conference Research on Computational Linguistics, pp. 19-33, 1997, Taiwan.

[12] Lin D., “An Information-Theoretic Definition of Similarity”, In Proceedings of the 15th International Conference on Machine Learning, pp 296-304, 1998.

[13] Mandala R., Takenobu T. and Hozumi T, “The Use of WordNet in Information Retrieval”, Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, pp 469--477, Montreal, CA, 1998.

[14] Miller G., Beckwith R., Fellbaum C., Gross D. and Miller K., "Introduction to WordNet: An On-line Lexical Database", International Journal of Lexicography, Vol 3, No 4, pp. 235-244, 1990.
[15] Niemann E. and Gurevych I., "The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet", Proceedings of the Ninth International Conference on Computational Semantics, pp. 205-214, 2011
[16] Platt J., " Fast Training of Support Vector Machines using Sequential Minimal Optimization", Advances in kernel methods, PP. 185 - 208, MIT Press Cambridge, ISBN:0-262-19416-3, 1999.
[17] Ponzetto S.,Strube M., "Knowledge Derived From Wikipedia For Computing Semantic Relatedness", Journal of Artificial Intelligence Research, Vol 30, No 1, pp. 181-212, 2007.
[18] Ponzetto S., Navigli R., "Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp 1522-1531, 2010.
[19] Resnik P., "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research, Vol 11, pp 95-130, 1999.
[20] Ruiz-Casado M.,Alfonseca E. and Castells P., "Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia", Proceedings of the 10th international conference on Natural Language Processing and Information Systems, pp 67-79, 2005.
[21] Suchanek F., Kasneci G., Weikum G., "YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia", Proceedings of the 16th International WWW conference, pp. 697-706, 2007.
[22] Toral a., Ferrandez O., Agirre E., Munoz R., "A study on Linking Wikipedia categories to WordNetSynsets using text similarity", In Proceedings of the international conference RANLP, 449–454, -2009.