International Journal of Computing Academic Research (IJCAR)
ISSN 2305-9184, Volume 5, Number 2 (April 2016), pp.122-132

http://www.meacse.org/ijcar

WIRS: Wisdom Based Information Retrieval System

Vijay Rana[1] and Vijay Dhir[2]
[1]I.K Gujral Punjab Technical University, Jalandhar, Punjab, India
[2]St. Solider Institute of Engineering & Technology, Jalandhar, Punjab, India

Abstract

Wisdom Information Retrieval System (WIRS) is proposed with the idea to give an interactive & a novel platform that shall execute tasks in parallel in discovering useful information and knowledge. In fact, WIRS is a knowledge based technique that numerically measures the amount of semantic similarity and relatedness between dissimilar words depending on the exploration of lexical resources such as WordNet. The proposed model shall prove to be a breakthrough in the domain of information retrieval making the information more meaningful. The existing work exploits mathematical and probability techniques to automatically extract valuable information from the web. WIRS has been implemented using NLP domain & the results obtained were analyzed using precision & recall method.

Keywords: Information Retrieval, Semantic Web, Knowledge Based, NLP.

Introduction

As the information on web is proliferating exponentially so are the existing information retrieval systems [26]. It is desired that these systems shall facilitate interoperability and also address the issues concerning semantic heterogeneity. For instance, S-Match [21] is semantic match making system that facilitates interoperability between different resources and performs element and structural level matching as well. PRIOR+ [9] enables semantic interpretability among different web application on semantic web and semantic heterogeneity has been addressed by few authors [12, 17, 24].In fact, achieving the interoperability between dissimilar information retrieval systems is extremely tedious, complex and error-prone task. Most of the existing search systems are not able to retrieve the desired results with their intended meaning and this is primarily due to the fact that these systems have not been designed with the intention of extracting wisdom from the web. Therefore, the need for research activities in information retrieval system supporting the heterogeneous infrastructure is apparent. To accomplish this vision this paper proposes Wisdom Information Retrieval System (WIRS) that shall push retrieval of meaningful results. WIRS executes in four autonomous phases where each phase has been implemented and tested individually as described later in this paper.

This paper has been broadly divided into five sections**.** Section 2 describes the work of eminent researchers highlighting the efforts being done to bridge the gaps in wisdom retrieval. Section 3 explains the proposed work and Section 4 discusses the results obtained. Section 5 finally concludes.

Related Work

WordNet [11] is an online lexical database based on psycholinguistic theories of human lexical memory. It assembles English words into sets of synonyms called synsets and manages the lexical information in terms of word meaning that defines the semantic relation between words. It contains more than 120,000 different word forms and 90,000 different word sense. A knowledge based system [12] that could handle the semantic heterogeneity by using semantic, name and statistical techniques was proposed by Maree and his team. The key idea behind this work is to find semantic correspondence between the entities

of inconsistence ontologies. Their work also enlightened the powerful role of semantics in ontology matching.

Recently [4, 14, 20, 22] the web and ontology communities have focused on classifying standards that are competent to handle heterogeneity problem on different environment. However this problem still remains because several earlier approaches are no longer competitive in dynamic environments. Authors [1, 2] have proposed to compute degree of semantic relatedness to retrieve the semantically related words. Gracia et.al in [7] have proposed web based semantic relatedness technique that numerically computes the degree of semantic relatedness between different ontology terms. The authors utilize Normalized Google Distance (NGD) [5] measure to compute the relatedness degree of co-occurrence of words on web pages. However, this approach still requires developing tools for dynamic configuration of new results into it.

The engraved study of literature indicates that much work are available in the domain but to the best of our knowledge and at the time of listing, very few works [21, 15] are available that have focused on designing a wisdom based information retrieval system. Hence the motivations for the current research work.

Wisdom Based Information Retrieval System (WIRS)

In contrast to the existing systems which focus on finding non-ambiguous words, WIRS focuses on initially finding the most ambiguous words and later computing the relatedness between the remaining query and the ambiguous word searched during initial phases. WIRS comprises of four modules namely Intelligent Context Selection Module (ICSM), Semantic Mining Module (SMM), Meaning Based Semantic Ontology Matching Module (MBSOM) and Query Matching Module (QMM).Figure 1 presents the abstract view of interaction of all four modules where ICSM gets activated when a user inputs a query and outputs the ambiguous keywords. The second module SMM identifies the set of maximum meanings for such ambiguous keywords which in turn are expressed as concise ontology terms with the help of MBSOM. These concise ontology terms are usually the Most Ambiguous Words (MAW)and QMM numerically calculates the degree of semantic similarity and relatedness between the MAW and the prominent words of the remaining query. Working of each module along with its algorithm is described in detail in the upcoming subsections.

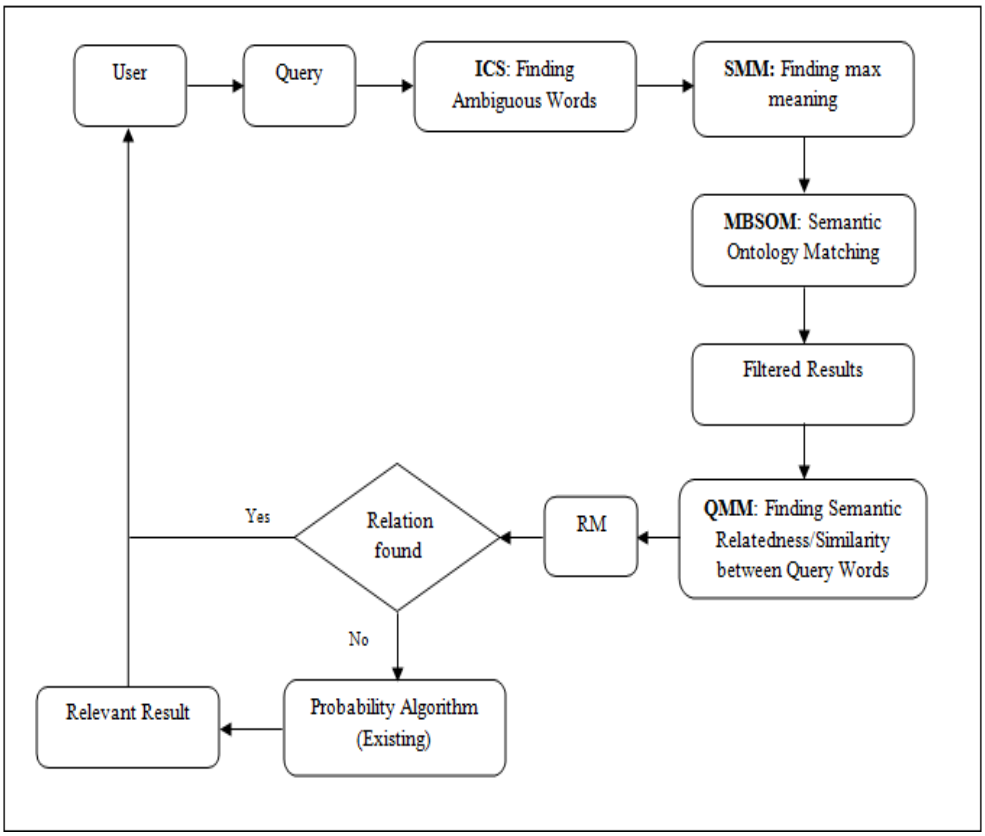


Figure 1: The Abstract View of WIRS

Intelligent Context Selection Module (ICSM)

A set of keywords is given as input to ICSM which finds out ambiguous keywords that describe the user's information needs. In contrast to the existing search engines that only retrieve the results on the basis of probable search while ignoring the semantics of user requirement, ICSM uniquely contributes a different & a novel algorithm that focuses on finding ambiguous queries which are further explored to find out the relevant meaning that describe the user desires. The algorithm tries to handle the ambiguity challenge which has been ignored by the existing algorithms [10, 19]. In fact, ICSM requires understanding the query structure and the associated semantics and on the basis of the structure thus defined, the rest of the modules (under WIRS) can automatically search the query corresponding to ambiguous keywords. ICSM is thus an intelligent module as it improves the probability of success by finding the appropriate results. Algorithm 1 describes the working of ICS module. The algorithm returns the most ambiguous word as illustrated with the help of case study presented later in this paper.

Algorithm 1: Intelligent Context Selection

Step 1; import WordNet, POS
Step 2: User Input t = "$Query$"
Step 3: Output = ambiguous keywords
Step 4: Tokenization text = $t.split("Query")$
Step 5: $pos.tag(text)$
$[(tk_1, WP), (tk_2, VB), (tk_3, AD), (tk_n, NN)]$
Step 6: Ignore the stop words
$stop = stopwords.words('text')/$
$/call\ stopword$
Step 7: Find ambiguous words in given query
$s = data\_input('text')$
$Txt = for\ s\ \in\ string.split()$
$for\ s\ \in\ Txt$
$if\ (len(WordNet.synset(s) > 1))$
$append\ (s)$
$end\ if$
$end\ for$
$end\ for$

Algorithm 1: The Intelligent Context Selection Module

Semantic Mining Module (SMM)

SMM involves defining the maximum possible meanings for the ambiguous words thus extracted by ICSM. SMM executes in two phases. It first identifies the set of related words and later determines the exact meaning of each occurrence. The first phase identifies lists of possible meaningful words in machine readable dictionaries [13] and other sense inventory systems [25]. It involves computing the similarity of the word contexts and utilizes external knowledge sources such as knowledge repository [8].The second phase pertains to mapping the context of an instance of the word to be unambiguous either with external lexical resources or with contents of earlier disambiguated illustrations of the word. SMM automatically identifies the set of possible meaning of each of the n words on the basis of synonym list attained from WordNet [11]. Algorithm 2 describes the working SMM. SMM uniquely mines the lexical information in terms of word meanings along with semantic relationship among the words. It also supports grammatical components that

having similar meaning & assemble them together to a form a single semantic entry such as synonym or synset, where a synset is a set of synonymous words.

Here W is the glossary is the set of words, those determined from WordNet Part of Speech (POS) [11] where POS is the set of hierarchy of parts of speech (n, a, v and r respectively). Hence SMM returns source synset which acts as input to next module. The implementation (given later) resulted into many redundant results and hence the limitation has been duly considered in MBSOM.

Algorithm 2: Semantic Mining Module

Input: Wordnet :wn; Ambiguous Keywords: txt

Output : Set of Integrated Senses: $\{s_1, s_2 \ldots\ldots s_n\}$

Step 1: import WordNet as wn

Step 2: Initialization of all word txt $= ICS()$;

$wn.synset('txt')$

$wn.synset.lemma()$

Step 3: Extraction and similarity computation

$if\ (wn.synste('txt') < 1)$

$then\ return\ ('txt')$

$else$

$for\ synset\ \in wn.synset('txt')$

$Sense = txt \in W \times POS \rightarrow 2^{SYNSETS}$

Step 4: Print the Words Relation

$echo\ Sense.hypemym()$

$echo\ Sense.hyponym()$

$echo\ Sense.synonym()$

$echo\ Sense.antonym()$

$echo\ Sense.troponym()$

$echo\ Sense.meronym()$

$end\ for$

Step 5: Set of integrated senses as output

$return(IS\{S_1, S_2, \ldots\ldots, S_n\});$

Algorithm 2: The Semantic Match Making Module

Meaning Based Semantic Ontology Matching (MBSOM)

Meaning Based Semantic Ontology Matching (MBSOM) module discovers correspondences among semantically related entities of ontologies and determines the set of synonym shaving different names and structures.

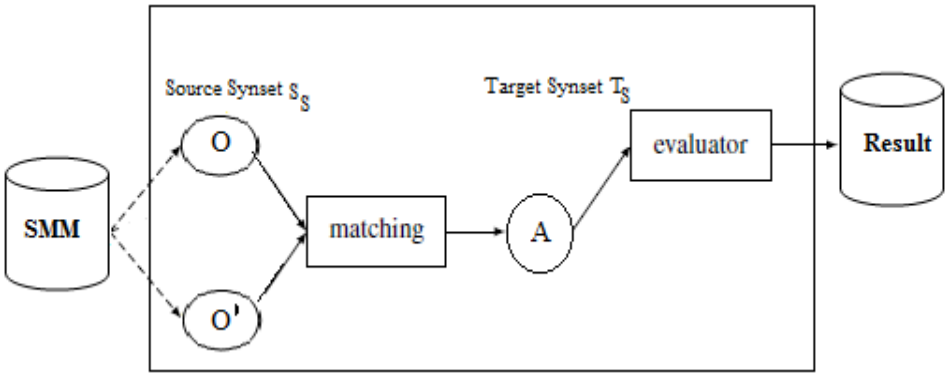


Figure 2: The Meaning Based Semantic Ontology Matching

Figure 2 highlights the working of MBSOM module. The output of SMM acts as the input to MBSOM and is referred as source synset (S_S). The ontologies are thus matched using the proposed MBSOM. The MBSOM retrieves information from the lexical information structure composing of a group of synonyms, antonyms and hypernyms words producing a target synset (T_S). Now, T_S is matched with S_S as per the algorithm 3.

Algorithm 3: Meaning Based Semantic Ontology Matching

Step 1: input S_S to MBSOM

$input: source\ synset: S_S$

$output: target\ synset: T_S$

Step 2: call MBSOM( )

2.1: compare S_S & T_S

2.1.1 $if\ \#\ of\ elements\ in\ S_S < \#\ no\ of\ elements\ in\ T_S$

$then\ S_S\ subsumes\ T_S$

2.1.2 $if\ \#\ of\ elements\ in\ S_S = \#\ of\ elements\ in\ T_S$

$then\ S_S\ equivalent\ T_S$

2.1.3 *else*

$S_S\ plugs\ in\ T_S$

Step 3: Calculate probability score & $find\ the\ words\ with\ highest\ probable\ score$

$$MAW = \sum_{P_S} Log(S_S) + Log(T_S)$$

Algorithm 3: Meaning Based Semantic Ontology Matching Module

The algorithm returns MAW i.e. the words with the highest probability score thereby reducing the redundancy. Next task is to establish the relation between MAW& the remaining query.

Query Mapping Module (QMM)

Query mapping module basically maps the entire query with the MAW thus generated through MBSOM. It finds the most appropriate meaning of ambiguous words according to the context in which it occur. Available literature [3,18] reflects that probability model & page rank algorithms have been used to resolve the issues relating to query mapping. Probability model is based on probability of relevant & non-relevant results while PageRank computes the back links of web pages. Both these algorithms neither

address the ambiguous queries nor do these compute the sentence and context related meanings of words thus found, therefore the motivation to propose QMM.

QMM computes the relationship between remaining query & the most ambiguous words on the basis of semantics & context. The relatedness measure between two or more words is computed either directly using the words in WordNet or the associated meanings of words those defined in WordNet respectively. The working algorithm of QMM is given in algorithm 4 and Figure 3 outlines the algorithm.

Algorithm 4: QMM

Step 1: input WordNet & $Part\ of\ Speech\ to\ module$

$RM = Relatedness\ Measure, QW = Query\ Word, MAW = most\ probable\ words$

Step 2: call POS( )and generate array of POS[]

/∗ compute relatedness ∗/

Step 3: if there are two MAW words only

3.1 $if\ (QW == null\ \&\&\ MAW == null)$

$then\ return\ (minSynset)$;

3.2: else

if(QW == MAW)

then call (MCI); /∗ $find\ most\ common\ information$ ∗/

3.3 MCI ($QW \in MAW$)

find IC factor using $pathFinder$( )

/

∗ find information content value between Query word synset & $most\ probable\ word\ synset$ ∗/

$RM =\ store\ highest\ IC\ value$

3.4 return (RM)

$endstep$ 3

Step 4: else/∗ $if\ there\ are\ more\ then\ two\ MAW\ words$ ∗/

/∗ $explore\ definitions\ of\ MAW_1, MAW_2, MAW_n\ from\ MBSOM\ module\ as\ D_n$ ∗/

4.1 call $Def.explore( ) = Definition[MAW_1.D_{n1}, MAW_2.D_{n2,} MAW_n.D_{nn}]$

/∗ $compute\ overlap\ between\ word\ definitions$ ∗/

4.2 $for\ each MAW_i.D_{ni} \in MAW_j.D_{nj}$

4.3 $RM = maxOverlap(MAW_i.D_{ni} == MAW_j.D_{nj})$

end for

returen (RM)

/∗ $Accordingly, most\ probable\ results\ (RM)\ are\ retreived$ & $QMM\ is\ resolved$ ∗/

Step 5: if RM ≤ τ && ∀≤ 0.5

$call\ prabability\ algorithm\ (PR)$

Step 6: return (MPPL) /∗ return most probable page links ∗/

Stop

Algorithm 4: Query Matching Module

Figure 3 outlines algorithm 4.

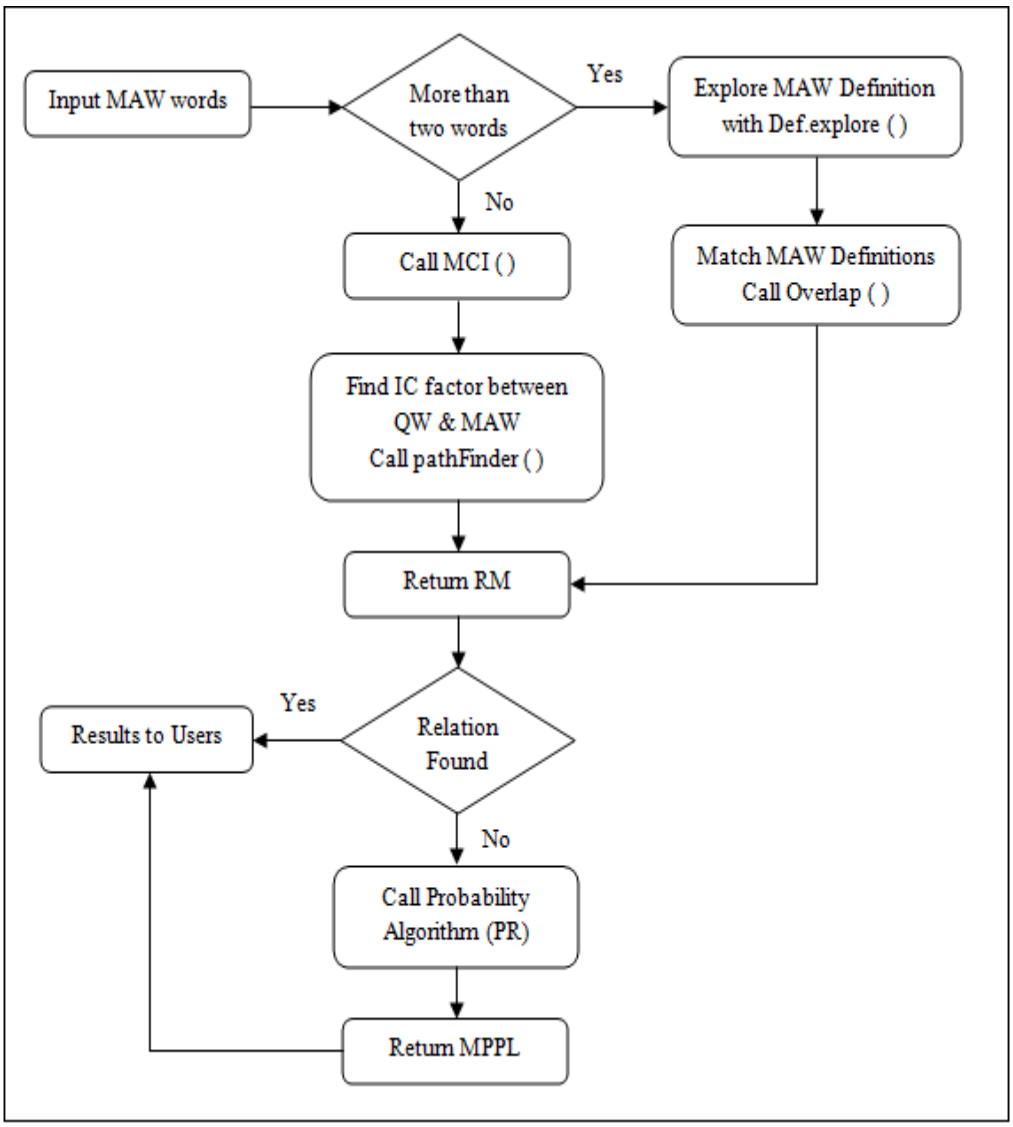
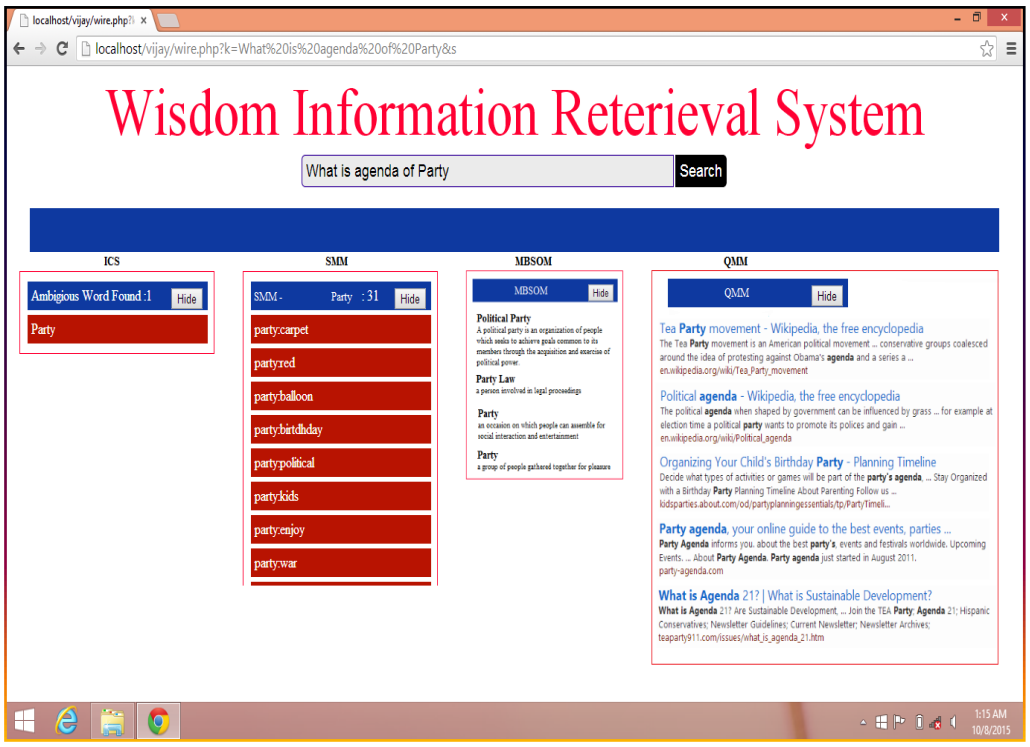


Figure 3: The Flow Diagram of QMM

The upcoming section describes the practical implementation of WIRS.

Implementation

WIRS has been designed & implemented on open source general-purpose scripting language PHP, including Apache, MySQL, PhpMyAdmin and Xdebug. The modules are being implemented with Knowledge based technique and integrated with WordNet 3.0, one of the popular databases for the NLP domain. We have used EditPlus as text editor and window 8 workstation with dual quad-core 3.00 GHz Intel Xeon processors. Figure 4 illustrates the results obtained on a query 'What is agenda of Party'.



Step-wise illustration

Step 1. Input user query: $Whatisagendaof Party$
Step 2. Execute ICS module:
Output: $Party$ $(Themostambiguous\ word)$
$Step$ 3. Execute SMM:
Output: 31 $associatedmeaningwithpartykeyword$ $(Relatedwords)$
Step 4. Execute MBSOM:
Output: 4 $mostprobablewords$ (MPW)
$Step$ 5. $ExecuteQMM$:
$Output$: $Agendaisrelatedtopoliticalparty$

Results & Discussion

The execution of QMM module could compute the relatedness measure with high precision. On input of different keywords, the relatedness measure could be calculated & a word with highest RM (Relatedness Measure) is considered to be the most desirable output. For example as shown in table 1, keyword Political party has the highest degree of RM & hence it is concluded that agenda is related to political party mostly. Word pairs show the two concepts and relatedness degree computed represents the similarity rate of each word pair ranging from 0 (no relatedness) to 4 (ideal relatedness).

Table 1: Semantic Similarity and Relatedness Measure

Words Pair		Relatedness Degree
Level 0	Level 1	
Agenda	Political Party	3.65
Agenda	Birthday Party	1.9
Agenda	Company	0.8

Further, in order to evaluate the overall performance of WIRS, we exploited the standard precision and recall to evaluate mapping results. These standard attempts to measure the amount of relevant and irrelevant information by evaluating the quantity of the obtained information. The overall performance is described in figure 5.Table 2 highlights list of 10 queries with their precision and recall measurer that describes the relevant and non-relevant results.

Table 2: List of Queries

Query No.	Quires	Number of link evaluated	More relevant	Less relevant	Irrelevant
Q.1	We sat along the bank of the Tevere river	5	4	1	0
Q.2	Book stays in London	5	3	2	0
Q.3	Total interest in last month	5	3	1	1
Q.4	They will definitely join our party	5	3	2	0
Q.5	Your flying planes with giant can be possible	5	2	1	2
Q.6	Apple	6	4	1	1
Q.7	Party	6	3	1	2
Q.8	Java	6	3	2	1
Q.9	Bank	6	5	1	0
Q.10	Interest	6	6	0	0

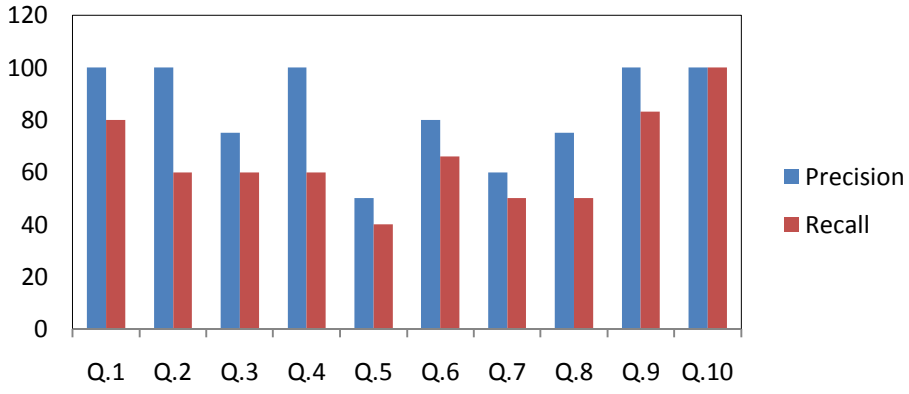


Figure 5: Describes the Precision and Recall Measure of Queries

Conclusion

WIRS is a novel system as it focused on finding the most ambiguous words and finding the relatedness measure with other important keywords in the query. It also considered removing redundancy among keywords being matched. The proposed modules are generic as these can be integrated with any platform and are application independent. A result obtained reflects that high precision and recall rate and thus proves the relevancy of WIRS.

References

[1] Agirre, Edmonds. P, Word Sense Disambiguation: Algorithms and Applications, Springer Publishing Company, Incorporated 2007, New York, pp 1-364, 2007.

[2] Alexander Budanitsky& Graeme Hirst, Evaluating WordNet-based measures of semantic distance. Computational Linguistics, vol. 32, no. 1, pp 13-47, 2006.

[3] AmitSinghal, Modern Information Retrieval: A Brief Overview, IEEE Computer Society Technical Committee on Data Engineering, pp 1-9, 2009.

[4] Ankita Kandpal, R. H. Goudar, Rashmi Chauhan, Shalini Garg and Kajal Joshi, Effective Ontology Alignment: An Approach for Resolving the Ontology Heterogeneity Problem for Semantic Information Retrieval,Advances in Intelligent Systems and Computing-Springer, pp 1077-1087, 2013.

[5] Cilibrasi, R.L., Vitanyi, P.M.: The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383, 2007.

[6] D. L. McGuinness and P. Pinheiro da Silva, Infrastructure for web explanations. In Proceedings of ISWC, pp 113–129, 2003.

[7] Jorge Gracia and Eduardo Mena, Web-Based Measure of Semantic Relatedness, WISE 2008, LNCS 5175, pp. 136–150, 2008.

[8] Lesk Michael, Automatic Sense Disambiguation using Machine Readable Dictionary: How to Tell a Pine Cone from an Ice Cream Cone, SIGDOC 86- ACM, pp 24-26, 1986.

[9] Mao Ming, Peng Y, Spring M, An Adaptive Ontology Matching Approach with Neural Network Based Constraint Satisfaction, Journal of Web Semantics, Volume 8, pp 14-25, 2010.

[10] Mark Sanderson, Ambiguous Queries: Test Collections Need More Sense, *SIGIR'08,* pp 16-24, 2008.

[11] Miller G.A, WordNet: A Lexical Database for English, COMMUNICATION OF THE ACM, Vol 38, No 11, pp 38-41, 1995.

[12] Mohammed Maree and Mohammed Belkhatir, Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies, Knowledge-Based Systems, Vol 73, pp199-211, 2015.

[13] Nancy Ide , Jean Veronis, MACHINE READABLE DICTIONARIES: WHAT HAVE WE LEARNED, WHERE DO WE GO?, pp 1-11, 1998.

[14] Ngamnij Arch-int and Somjit Arch-int, Semantic Ontology Mapping for Interoperability of Learning Resource Systems using a rule-based reasoning approach, Expert Systems with Applications, Volume 40, Issue 18-Elesivier, pp 7428–7443, 2013.

[15] P Shvaiko, J Euzenat, Ontology matching: state of the art and future challenges, Knowledge and Data Engineering, IEEE, pp 158-176, 2013.

[16] P. Pinheiro da Silva, D. L. McGuinness, and R. Fikes, A proof markup language for semantic web services. Technical report, KSL, Stanford University, 327-348, 2004.

[17] Pinto, Carlos Sousa, Jayadianti, Herlina, Nugroho, Lukito Edi and Santosa, Paulus Insap, Solving problems of data heterogeneity, semantic heterogeneity and data inequality : an approach using ontologies, Journal CAlg - Artigosemlivrosdeatas, pp 11-24, 2012.

[18] R. Armstrong, D. Freitag, T. Joachims, and T.Mitchell, Webwatcher: A learning apprentice for the world wide web. In Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, 1995.

[19] R.K. Roul and S.K. Sahay, An Effective Information Retrieval for Ambiguous Query, AJCSIT, Vol. 2, No. 3, pp 26-30, 2012.

[20] Shrutilipi Bhattacharjee and Soumya K. Ghosh, Measurement of Semantic Similarity: A Concept Hierarchy Based Approach, International Conference on Advanced Computing, Networking and Informatics, pp 407-416, 2015.

[21] Shvaiko P, F. Giunchiglia, Web Explanations for Semantic Heterogeneity Discovery, ESWC 2005, LNCS 3532, pp 303-317, 2005.

[22] Simon Scheider and Werner Kuhn, How to Talk to Each Other via Computers: Semantic Interoperability as Conceptual Imitation, Applications of Conceptual SpacesVolume 359 of the series Synthese Library, pp 97-122, 2015.
[23] Vijay Rana, Blueprint of an Ant-Based Control of Semantic Web, International Journal of Advancements in Technology (IJoAT), pp.603-612, 2011.
[24] Vijay Rana, Singh G, "An Analysis of Semantic Heterogeneity Issues and their Countermeasures Prevailing in Semantic Web", ICROIT 2014, DOI: 978-1-4799-2995-5, IEEE Xplore, pp 16-22, 2014.
[25] www.isical.ac.in/~lru/wordnetnew/index.php/site/index.
[26] Y Gupta, A Saini, AK Saxena , A new fuzzy logic based ranking function for efficient information retrieval system, Expert Systems with Applications, Elsevier, pp 79-86, 2015.