# A Synonymous Approach to Enhance Relevance in Search Engine

Dr.V.Gomathi* E. Padma#, I.Piriya Dharshni#

**Abstract**— Web pages are growing at a galloping rate. Presently there exist 300+ million web pages. When a user provides search keywords in the existing search engines, the keyword is searched for its match in the titles, headings, or special fields called meta-tags and the first few paragraphs of the web page and displayed. But, if the meta-tags are framed wrongly either intentionally or unintentionally, this word matching technique may lead to an irrelevant page. To alleviate the above mentioned problem, this project aims to provide an efficient way of searching the web contents by proposing "Synsearch" - a synonymous approach based middle-ware. The novel Synsearch gives a clear separation between web content, presentation and the meaning of the content. This is more useful to make the web pages as machine processable and thus enhances relevance in search engine. Synsearch will unburden the users from interpreting results of search engines. The proposed method alters the ranking algorithm used in existing search engines based on probability of word occurrences. Synsearch further alters the searching mechanism by introducing a synonymous approach.

**Index Terms**— Relevancy, Search Engines, Semantic Web, Stop-words, SynSearch, SynSets, WordNet.

## 1 Introduction

The World Wide Web (abbreviated as WWW or W3, commonly known as the Web), was developed by Sir Tim Berners-Lee. Web is a system of interlinked hypertext documents accessed via the Internet. Web is one of the services that run on the Internet. First generation of web consists of web pages inter mixed with content and presentation. A change in either one will affect the other. The second generation made a separation between content and presentation which was made by using style sheets in HTML. The upcoming third generation targets on separating between content, presentation and semantics. This project focuses mainly on the third generation of web.

Syntactic Web is a phrase meant to describe the current, mostly HTML-based World Wide Web. The term stems from the root word "syntax", which is the mechanics of a language used to convey information. A syntactic web page is any document on the web that does not contain special tagging to convey meaning. It is difficult to parse the meaning by a computer program. The computers are only used to display the information, that is, to decode the colour schema, headers, and corresponding links encoded in web pages. The interpretation and identification of relevant information is delegated to human beings. Of course, the interpretation process is very demanding and requires great effort to evaluate, classify and select relevant information. Because, the volume of available digital data is growing at an exponential rate, it is becoming virtually impossible for human beings to manage the complexity and volume of the available information. This phenomenon is often referred to as "information overload" and it poses a serious threat to the very usefulness of today's web. Here the size of results available is large. This situation gets really worse as the size of the web increases. Most users only browse through the top results, discarding the remaining ones. A web search engine is a software tool that is designed to search for information on the World Wide Web.

## 2 Existing Works

Search engines are very important tools for the people to get information from Internet but the low-accuracy and low-recall persists widely in current search engines. There are several methodologies that try to find lexicalizations for a keyword like Latent Semantic Analysis. It considers that words that co-occur in the same documents are semantically related. However, it tends to return domain related words but, sometimes, not truly "equivalent" ones.

David Sanchez et al, proposed, an automatic and autonomous methodologies for semantic disambiguation and classification of web resources and for discovery of lexicalizations and synonyms for a specific domain [1]. Smart Web Query Engine project [2] investigates a new web query method that is called Smart Web Query (SWQ) method for semantic retrieval of web data. Paolucci et al[3,4] developed a Matching Engine that performs flexible matches recognizing the similarity of an advertisement in UDDI and a user request.

Automatically discovering synonyms from large corpora and dictionaries has been popular topics in natural language processing [5,6]. Ontology with the techniques of natural language processing is used for document indexing and ease of search [7]. Hogan et al [8] suggested an object-orientated model to firstly integrate data about the same object from multiple sources, and secondly enable expressive queries over the integrated information space. Tim et al [9] developed "Swoogle", a crawler-based indexing and retrieval system for the Semantic Web documents – i.e., RDF or OWL documents that are indexed by character N-Gram or URIs.

Search results might contain a large number of entries, but they might not have high recall and precision rates. For example, a search for web pages where "TCP/IP" and "protocol" occur might return all relevant web pages, but the result would be of very little use if the user had to sort through through 22,500,000 web pages of little interest. Also, search results are sensitive to the vocabulary used. Indeed, users frequently formulate their search in a vocabulary different from that which the relevant web pages adopt. In the TCP/IP example, the relevant web pages might use "standard", instead of "protocol"; hence the web pages would

not be the best match for a search using the keywords "TCP/IP" and "protocol".

## 3 PROPOSED SYSTEM

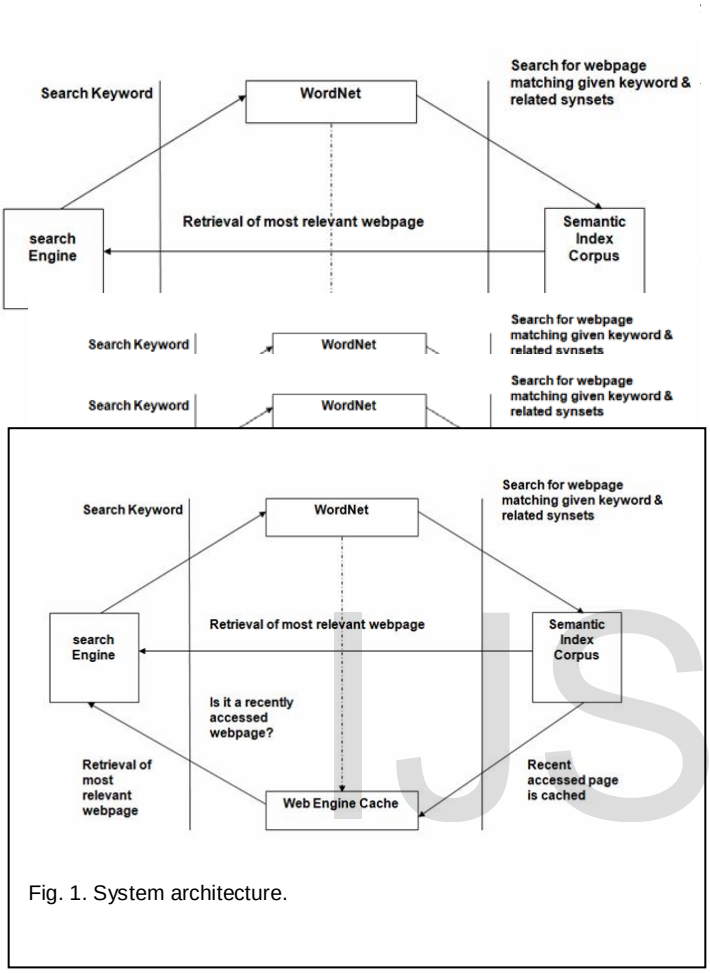


Fig. 1. System architecture.

The "Semantic Index Corpus" consists of the web page link with the relevant meta-tags that are framed by semantic web developer. When the user enters the search query in the search engine query box, the system sends the word to WordNet and produces all the possible meanings that can lead to that query and redirects the search to the database. So, now the most relevant web page is returned by the database. If the web page was most recently accessed then the web page will be returned from the web engine cache. This further reduces the search time.

The proposed system is modularized into two phases. The first phase forms the semantic web development and the second forms the search algorithm development. System design represents the overall design of the proposed system "synsearch". This includes the details about the different modules in the system along with the various diagrams that represent the various activities related to the system.

This phase-I includes the procedure that has to be done offline. This work is carried down by the web developer and the user is blind about it. This web development phase has the following stages:

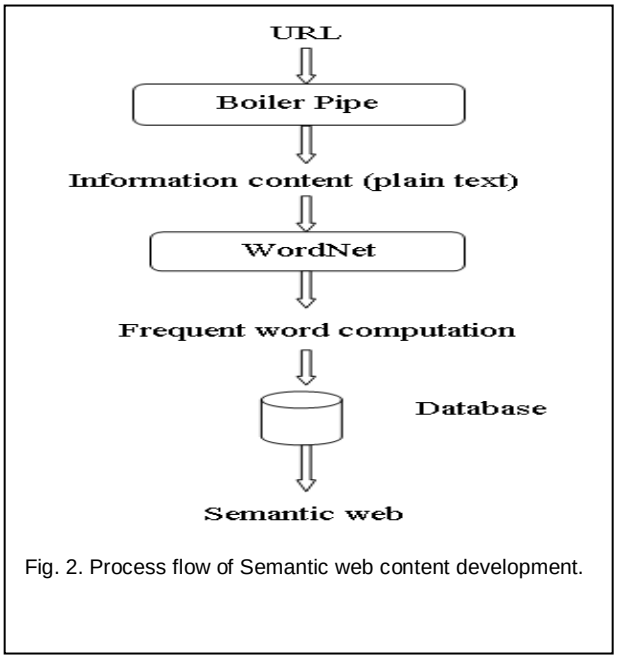


Fig. 2. Process flow of Semantic web content development.

### 3.1 Web Content Extraction

The information content of the webpage is extracted by passing the URL as input to Boilerpipe which is a API. This API contains many classes which can be used to extract various formats of content from the given web page URL. Among these classes, the proposed synsearch uses "ArticleExtractor" which extracts the plain text content in the web page. This is done by detecting and removing the surplus "clutter" that includes images (in any format), rendering details (font size, font colour,.. etc), references, hyperlinks and so on. The function getText() for which the url is passed as an input will return the extracted content.

### 3.2 Stop-words Removal

In general, the English sentences have the presence of almost about 60% or even more than that of stop-words (is, a, the, an, usually, obviously, etc). Since these words convey no meaning, their removal will not affect the semantics of the document, whereas it increases the system processing time. This module converts the document into more interpretable format. The extracted content is tokenised into separate words and they are compared with the pre-stored stop-words list in the system and are eliminated for further processing.

### 3.3 Synonymous Approach

WordNet is an electronic lexical network developed since 1985 at the Princeton University by a linguists and psycholinguists team of the Cognitive Science Laboratory. The advantage of WordNet is the diversity of the information that it contains (large coverage of the English language, definition of each meaning, sets of synonyms and various semantic relations). In addition, WordNet is freely usable.

WordNet is a large lexical database with synsets (sets of synonymous terms) for a given word. The synsets contains

collection of more words which is synonymous to the given word. Depending upon the vagueness of the word, there will more than one synsets for the given word. The words from extracted content are obtained and stop-words are removed. The remaining words are sent into the WordNet. This returns the associated synset/synsets. These synsets are tokenized into separate words and similar words are filtered.

For example, if there are words like 'computer' and 'computers' or 'reckoner' in the document, then after stop-words removal and wordnet process, these are identified to be synonymous word and they can be replaced with the word 'computer'. Thus clumping of synonymous terms is performed.

### 3.4 Frequent Word Computation (Ranking)

In this module the number of occurrences of the word is computed. If the word occurs for the first time then its occurrence is initiated for the respective word and when the same word appears again in the document its occurrence value is increased. This process is repeated until the document ends. It provides the number of occurrences of the word and using which the probability of a particular word can be computed using the formula given below:

$$Pwi = \frac{Cwi}{\sum_{i=1}^{N} Cwi} \quad (1)$$

where, Pwi is the probability of the ith word

Cwi is the number of occurrences of the ith word

N represents the total number of words in the document after removing stop-words and clumping synonymous terms.

After sending the words to WordNet and clumping the synonymous terms the total number of words has been reduced whereas, probability of words has increased. The words which are having higher probability are given higher priority which will be considered in computing the probability density function (pdf).

$$pdfdoc = \sum_{j=1}^{k} Pwj \quad + \sum_{j=k+1}^{n} Pwj \quad (2)$$

If pdfdoc < α, then k is incremented by 1, where α is the threshold frequency (α ={0..1}). The threshold frequency is used to limit the words in the web document. For a sample document with 6035 words, the threshold frequency is set with 0.3, 0.4, 0.5, 0.6 and 0.7 for both the syntactic approach and the synonymous approach.

TABLE 1
THRESHOLD FOR SYNTACTIC APPROACH

| α | No. of Priority Words |
|---|---|
| 0.3 | 50 |
| 0.4 | 97 |
| 0.5 | 169 |
| 0.6 | 272 |
| 0.7 | 425 |

TABLE 2
THRESHOLD FOR SEMANTIC APPROACH

| α | No. of Priority Words |
|---|---|
| 0.3 | 12 |
| 0.4 | 28 |
| 0.5 | 53 |
| 0.6 | 95 |
| 0.7 | 165 |

Tables 1 and 2 clearly show that, the higher threshold frequency (α) leads to large number of priority words, which contains many low frequency words also. Thus, here the threshold value for probability density function is limited to 0.5 with a hypothesis that those words are the most relevant to document. With this computation the offline phase of the synsearch is done.

For example, a sample document with probability value for few words before sending to WordNet is given in Table 3.

TABLE 3
PROBABILITY OF WORDS IN SAMPLE DOCUMENT BEFORE SENDING TO WORDNET

| WORD | COUNT | PROBABILITY |
|---|---|---|
| computer | 93 | 0.030412 |
| computers | 60 | 0.019621 |
| Program | 38 | 0.012426 |
| Memory | 34 | 0.011118 |
| Machine | 27 | 0.008829 |
| instructions | 26 | 0.008502 |
| programs | 23 | 0.007521 |
| Main | 21 | 0.006867 |
| Called | 19 | 0.006213 |

After clumping the synonymous terms the probability value is given in table 4.

TABLE 4
PROBABILITY OF WORDS IN SAMPLE DOCUMENT AFTER SENDING TO WORDNET

| WORD | COUNT | PROBABILITY |
|---|---|---|
| Computer | 162 | 0.045412 |
| Numbers | 142 | 0.005559 |
| Program | 105 | 0.012426 |
| Read | 105 | 0.003924 |
| Instructions | 77 | 0.008502 |
| Memory | 59 | 0.011118 |
| Called | 54 | 0.006213 |
| Perform | 53 | 0.003597 |
| Jump | 45 | 0.002616 |
| Machine | 33 | 0.008829 |

The word occurrence counts before and after synonymous approach is depicted in figure 3 and figure 4.

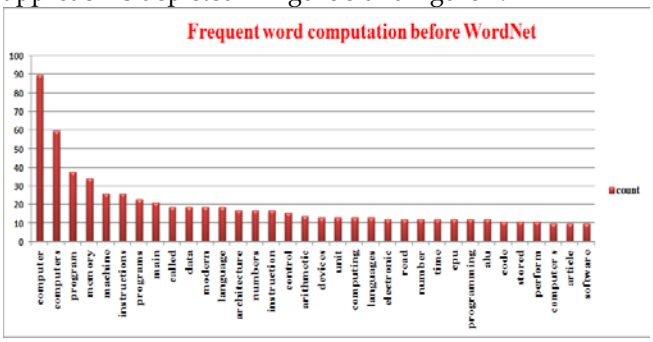


Fig. 3. Word count for a sample document before sending to WordNet

Due to clumping of the synonymous terms, there are only unique words found in figure 4.

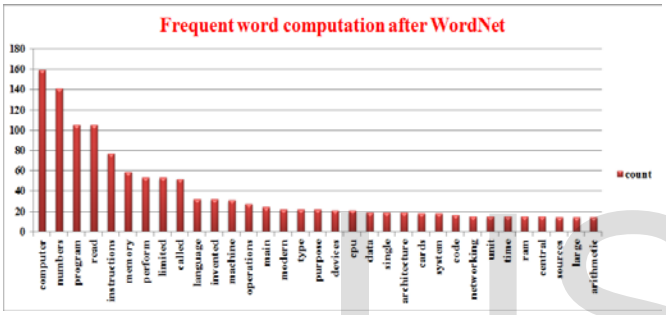


Fig. 4. Word count for a sample document after sending to WordNet

## 4 Search Algorithms

A search algorithm is an important phase in developing a search engine. In this proposed system a synonymous approach is followed which offers better relevant results. When the user types his search query in the search box it is tokenized and passed into the wordnet to obtain its synsets. Then these synsets are matched with database "Semantic Index Corpus" and using the ranking scheme the results are arranged accordingly. The ranking methodology is by using probability of frequent word occurrences.

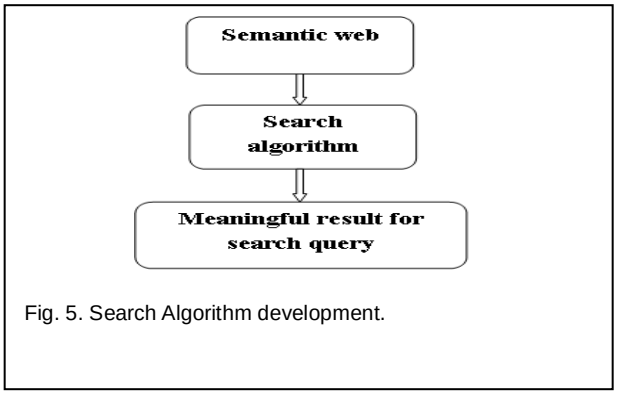


Fig. 5. Search Algorithm development.

## 5 System Implementation And Comparative Analysis

The proposed system is implemented in Java Platform. The Boilerpipe and WordNet Tools are much involved in phase-I implementation.

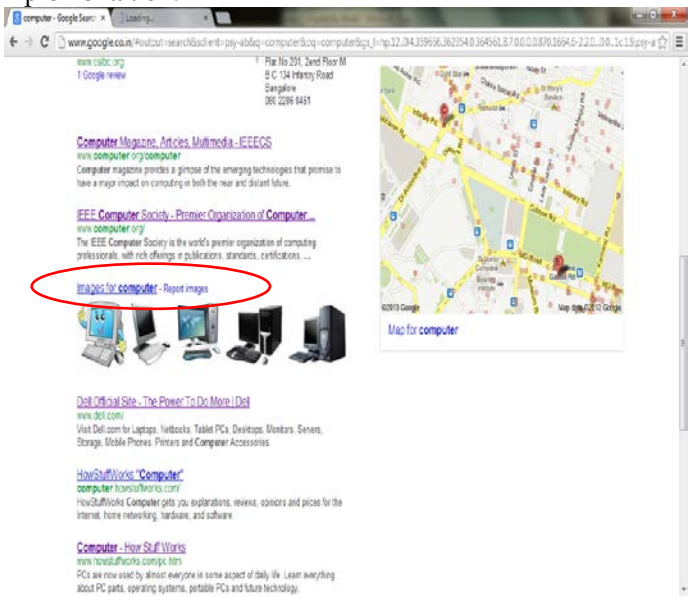


Fig. 6. Conventional search engine result page for keyword 'computer'

The above result page as shown in Figure 6 is produced by google for the keyword "computer". The source document for the encircled link is given in Figure 7.

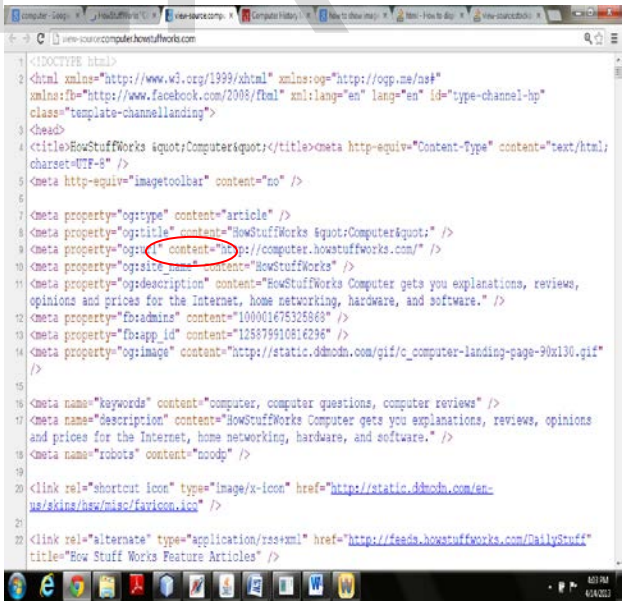

Fig. 7. Source page of encircled link

The content of the link is given below:

```
Internet
We look to the Internet for news, socializing, shopping, research and more. From HTML code to instant mes-saging, we'll break down what's re-ally going on whenever you log on, send an e-mail, visit a popular Web site or post to a blog.
```

The web page of the above extracted content is shown in Figure 8.

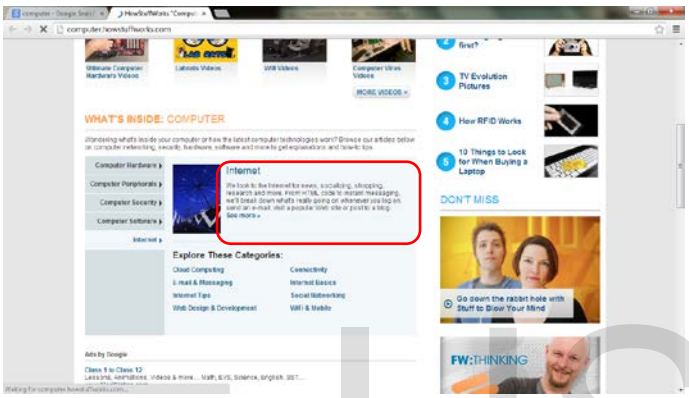


Fig. 8. Browser view of encircled link

Here, the meta-tag of source file contains the word computer and hence it is produced as result even it does not contain the relevant content.

But, the proposed system follows ranking methodology by using probability of word occurrences. This proposed ranking method eliminates the above said link because; it does not contain the word computer frequently. The marked link gets the rank 8 in the proposed ranking mechanism.

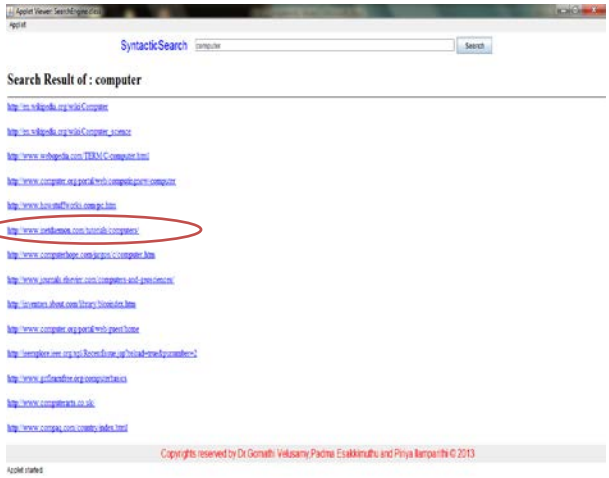


Fig. 9. Result page of proposed syntactic approach

This paper further analyses the proposed synonymous approach in search mechanism. Thus only the web pages with relevant contents are displayed as result. Here the marked link gets the rank 6. Thus the relevancy in enhanced by the proposed system.

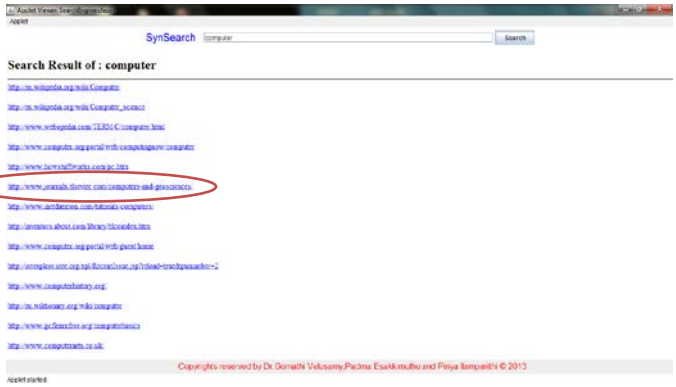


Fig. 10. Result page after enhancing relevancy in search algorithm

The user can view the web page by clicking on the link displayed in the Search Engine Result Page (SERP). The Figure 11 shows the web page displayed as a result of clicking the first link on the SERP.

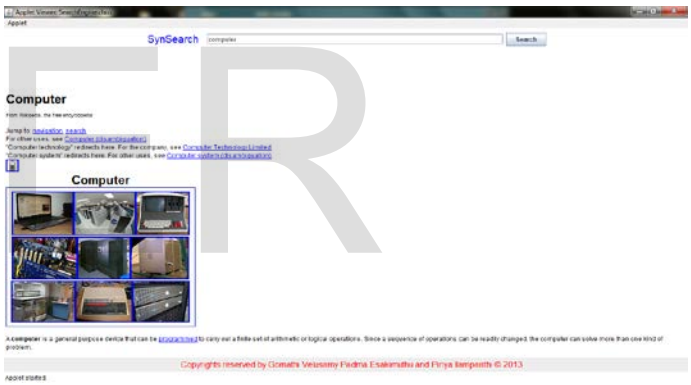


Fig. 11. Web page display by clicking link on SERP

## 6 Conclusion

The proposed synsearch saves the user's precious time and reduces the reading overhead of the user. Also, this does not provide any new technique for the user to learn. It simplifies the user's process without even the knowledge of user. This system provides enhancement in both search algorithm and the ranking mechanism. Thus it provides the relevant result for the web page and also they are arranged according to the ranking mechanism. This system reduces the size of the search result in search engine and also it eliminates the problem of "information overload". This system uses the synonymous approach which is a part of the semantic web. In future, the language grammar can be added to the synonymous approach so that if forms a complete semantic web and search engine. The misspelled words can be corrected by suggesting similar words as provided in existing search engines like Google.

## REFERENCES

[1] David Sanchez and Antonio Moreno, “Development of new techniques to improve Web search”, in Proc. of 9th International Joint Conference on Artificial Intelligence, 2005.
[2] Roger H.L. Chiang, Cecil Eng Huang Chua and Veda C. Storey, A smart web query method for semantic retrieval of web data, Data & Knowledge Engineering, Volume 38, Issue 1, pp.63-84, 2001.
[3] K. Sycara, M. Paolucci, A. Ankolekar and N. Srinivasan, "Automated Discovery, Interaction and Composition of Semantic Web services," in Journal of Web Semantics, Volume 1, Issue 1, pp. 27-46,2003.
[4] N. Srinivasan, M. Paolucci, and K. Sycara, “Adding OWL-S to UDDI, implementation and throughput”, in First International Workshop on Semantic Web Services and Web Process Composition, California, USA, 2004.
[5] Sanchez, D. and A. Moreno, “Automatic Discovery of Synonyms and Lexicalizations from the Web”, In Proceedings of the 8th Catalan Conference on Artificial Intelligence, 2005.
[6] Bollegala, D., Y. Matsuo, and M. Ishizuka, “Measuring Semantic Similarity between Words using Web Search Engines”, in Proceedings of the 16th international conference on World Wide Web (WWW), 2007.
[7] A. Bouramoul, M-K. Kholladi, B-L. Doan, “How Ontology Can be Used to Improve Semantic Information Retrieval: The AnimSe Finder Tool”, In International Journal of Computer Applications, Vol.21, No.9,pp.48-54,2011.
[8] A. Hogan et al., “Searching and Browsing Linked Data with SWSE: The Semantic Web Search Engine,” Journal of Web Semantics, 2011.
[9] Tim Finin, Yun Peng, R. Scott, Cost Joel, Sachs Anupam Joshi, Pavan Reddivari, Rong Pan, Vishal Doshi and Li Ding, “Swoogle: A search and metadata engine for the semantic web”, In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, 2004.
10]Princeton’s WordNet site: http:// www. cogsci.princeton.edu/~wn/
[11]http://www.readwriteweb.com/archives/semantic_web_authoring_tools.php
[12] http://code.google.com/p/boilerpipe