ISSN 2320 - 2602
Volume 3, No.5, May 2014
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2014/ijacst05352014.pdf

Improvement of Web Search Results using Genetic Algorithm on Word Sense Disambiguation

Pooja Bassin[1], Manisha[2]
[1]M.Tech Scholar, India, pooja.b2809@gmail.com
[2]Associate Professor, India, mani_bhardwaj@yahoo.com

ABSTRACT

Extraction, finding content on the Internet using web search engine outputs web pages in huge amounts. Retrieval of information is a tedious task for extracting some meaning information over the web where the web is in itself a network of networks. The information stored on the web is spread across various platforms. There is no centralization of information on the Internet. The results that are displayed to the user are based on the keywords or based on their synonyms. The web pages or documents displayed to the user are not the best or reaching up to the level of user's expectations. Hence it becomes the user's task to filter out relevant information out of the huge amount of pages. This is one of the major research issues in Information Retrieval. This problem can be sought out with the help of mining of data on the basis of intention of the user. Depending on the intention of the user while entering the query on the search engines, information should be displayed. Such kind of technique would help in user satisfaction at great level.

Keywords: Intent Mining, Word Sense Disambiguation, Genetic Algorithms, Web Search, Intelligent Agent

1. INTRODUCTION

Information retrieval is the foundation for modern search engines. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing [1]. Retrieval of content or information on the Internet has been a popular and fascinating task, it is indeed till date but has become more tedious and day by day due to the advancements of new technologies it is expected from the web search engines as well that they improve their searching techniques so as to return more relevant information to the user rather than just returning web pages and annotations containing only the keywords entered by the user in his or her query. An intelligent agent is required which can help the search engines in retrieving the information which is more meaningful and relevant for the user depending on his current interests, personalized information, searching history or any other feature. This would not only improve the search results but also save user's time in filtering out the relevant links based on the query entered by the user.

2. WEB SEARCH

In today's era, anything and almost everything is available on the Internet. People search the web for information that is stored across various platforms on numerous websites. People nowadays search for information which they want to acquire for their knowledge. The information is scattered all around the web for which no particular agency or organization is responsible for its content and availability. The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories [2]. Hence users rely on the search engines for best output of their queries. The search engines are responsible for returning the best link for a user's query that satisfies his interests. On the other hand, the search engines also depend on the user's query to be more and more precise, structured and with the most essential and relevant keywords which are required to be searched over the entire Internet. The better the query the relevant the result set. Today, ample of search engines are available of the most famous software companies which claim to return the optimized results for the user's query. Google, Bing, MSN, Yahoo are some of the examples of the most prevalent and popular search engines till date.

3. WORD SENSE DISAMBIGUATION

Word sense Disambiguation, as the name suggests that there is ambiguity in the sense of a word. On further elaborating it can be understood as the word can have more than one meaning in different contexts [3], [4]. For example we can look at the following sentences:

a) John works at the *stock* market.

b) You need to add some vegetable *stock* before preparing the soup.

The word *stock* in the above mentioned sentences has absolutely two very different meanings because it refers to completely opposite scenarios. In the first sentence, *stock* represent to business, market or shares. Whereas, in the second sentence *stock* relates to something regarding food. In both the sentences, same word with same spelling is used but has totally different meanings. The meanings are understood based on the frame of reference in which they have been used. To comprehend the correct meaning of an ambiguous word is a crucial task for any computerized system. Hence our system tries to work in this area for better understanding of the meaning of the ambiguous words looked for on the Internet.

4. GENETIC ALGORITHM

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems [5].Genetic Algorithms belong to the category of Evolutionary Algorithms. Genetic Algorithms are being studied extensively for web searching so as to optimize the results given by the search engines. Genetic Algorithms copy the real life scenario of biological life of living beings. In real world, the birth of a child is the result of parents reproducing and the operations of crossover and mutation being performed on the child's genetic structure that help in constituting his or her overall structure, appearance, features etc. Similarly, in computer programming this real world scenario is replicated on computer programs so as to achieve optimized results. During our study we have seen that immense work is being done in the field of Genetic Algorithms and these are also being used for web searching.

5. EXPERIMENTAL SETUP

In this study we have tried to apply genetic algorithm for searching relevant content on the Internet. Ambiguous words of English language such as bass, orange, bank, crane etc have been taken into consideration. The system has been developed on Java platform.

6. METHODOLOGY

Figure 1 shows the methodology of the system. The user enters the query on the search engine. Based on the query entered by the user a set of population of individuals in binary encoded form are created. These individuals are passed to the genetic algorithm. Selection, crossover and mutation operators are applied and a new population is generated. Weighted sum of the individuals in the new population are calculated and the maximum weight is retrieved. It is then matched with the threshold value and hence sense is deduced from the database. A refined query is formed with the sense identified and the results are sent back to the user.

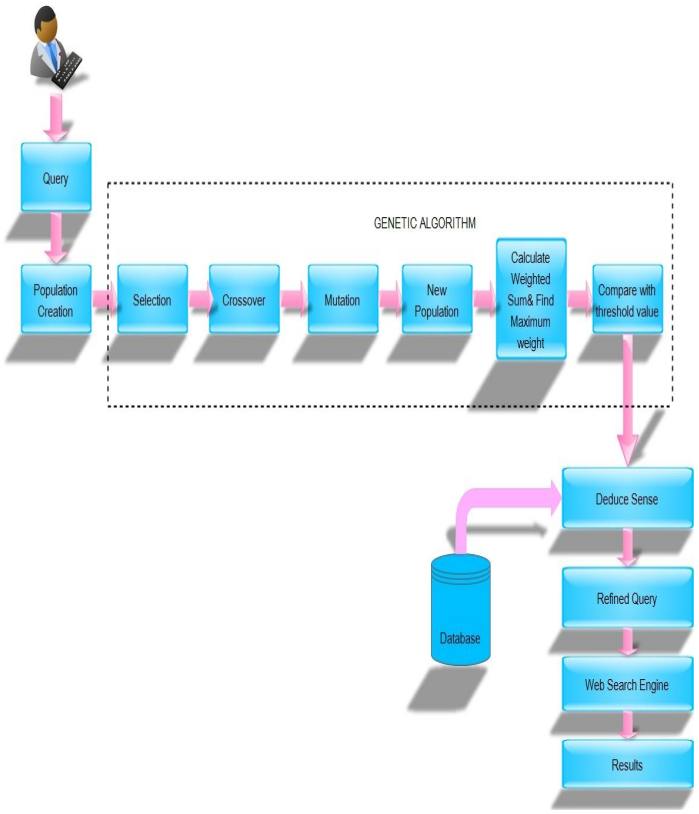

Figure 1: Genetic Algorithm Based Word Sense Disambiguation Framework

6.1 Representation of an Individual

In our study, the individual is represented with five genes distinguished by the five prominent attributes of the user's profile and queries. His profession, hobbies, additional interests, query and its meaning. Figure 2 shows how the individual has been represented in our study:

a_1	a_2	a_3	a_4	a_5

Figure 2: Representation of an individual

6.2 Fitness Function

The fitness function or the objective function of a Genetic Algorithm is the criteria for determining an individual's

value for retaining itself in the population for further operations or not. It can be understood by a simple example. For example we have four individuals:

101101
111000
000001
100010

In the above mentioned individuals the fitness function can be taken as the maximum number of 1s. So the fitness function for each individual is 4, 3, 1, and 2. Hence the individuals with maximum fitness value are retained for operations on itself. Rest of the individuals is discarded from the population.
The fitness function for our system is the precision for all the positive attributes retrieved by set of all the attributes.

6.3 Selection Operator

The selection procedure chosen is Tournament selection where each individual is matched with other and the one with higher fitness value is chosen as the winner to carry out further operations. In the above example of fitness function the individuals are compared randomly with each other, say, first and second and third and fourth are compared. So in first case individual one is the winner as it has higher fitness value and in the second case individual four is the winner. These two winners are chosen for crossover and hence become parents of their upcoming off springs.

6.4 Crossover Operator

The very next operator immediately to be applied after selection is crossover. The two individuals who are declared as winners in the Tournament selection go through crossover to generate two new children. The crossover site is chosen to be 2 in our study and the type of crossover being used is one-point crossover. Figure 3 shows an example of One-Point Crossover:

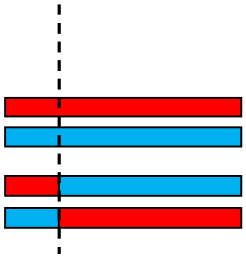

Figure 3: One-Point Crossover

The children born out of crossover operation replace the old population thus forming the new population. The crossover probability in our study is kept to be 1 i.e. P_c=1.

6.5 Mutation Operator

The next operator that comes after crossover is mutation. In our study the type of mutation that is studied is flip bit mutation. One of the individuals is chosen at random from the new population for mutation. Each individual is previously encoded in binary format hence one of the genes of the individual is chosen at random. The gene randomly chosen is mutated i.e. the bit is changed from 0 to 1 or from 1 to 0. This new individual after mutation is added back to the new population after crossover hence forming the final population and completing one generation of Genetic Algorithm.

6.6 Weighted Sum

After successful execution of one generation and formation of new and final population of individuals the weighted sum of each individual is calculated. The Weighted Sum or Sum of weights, S_w, is calculated using the following formula:

$$S_w = \sum_{i=0}^{5} w_i a_i$$

where,

w_i = weights.
a_i = attributes.
Out of all of the weighted sums calculated by the above mentioned formula the weight with maximum value is retrieved and matched with the threshold value. Respective sense is deduced from the Lexicon depending on the threshold value. In our study we have kept the threshold value as 0.60. Hence if the maximum weight is greater than 0.60 then sense one is taken under consideration else sense two. But if the maximum weight comes same as the threshold value then both the senses are shown to the user and hence it is left up to the user to decide which meaning to go for.

6.7 Refined Query

After deducting the correct sense from the Lexicon a reformulated or refined query is generated by appending the sense deducted and the query entered by the user. The refined query is redirected to the search engine and results are displayed to the user.

7. GENETIC ALGORITHM BASED WORD SENSE DISAMBIGUATION ALGORITHM

Table 1 shows whole of the above procedure summarized in the following algorithm:

Table 1. Algorithm for GA-WSD

Input:
Query entered by the user.
User's registered details.
Process:
Preprocess the query entered by the user.
Generate all the possible combinations.
Individuals are created from all the combinations.
Perform following to find maximum weight individual by applying Genetic Algorithm.
Generate initial population by passing the individuals to GA.
Do for one generation.
Calculate fitness function for each individual in the population.
Apply Tournament selection operator for choosing winners to fight by calculating each individual's fitness.
Let two and two fight.
The winners make children after performing crossover.
Children make new population.
Randomly select an individual for mutation.
Mutate bit from 0 to 1 or 1 to 0 by selecting random position of the individual.
Final population is created after including the mutated individual.
Calculate weighted sum of each individual in the population.
Find maximum weight individual.
If maximum weight is greater than the threshold value deduces first sense from the database else deduce second sense.
Output:
Refined query generated.

8. EVALUATION AND RESULTS

In Information Retrieval the two most important measures for evaluating any system's performance are precision and recall. In our study, we have considered only precision for evaluating our system's performance. Recall has not been taken under consideration here. The above system is evaluated on the basis of precision. The formula for precision is:

$$\text{Precision} = \frac{|\{\text{relevant documents}\}| \cap |\{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Table 2 shows a comparative table of precision results which depicts the GA-WSD results of our study and on the other hand compares it with the conventional search engine's results.

Table 2. Precision table for GA-WSD

Query	Cases	GA-WSD Results		Google Results	
		Page 1	**Page 2**	**Page 1**	**Page 2**
Bank	C1	0.93	1.00	0.00	0.00
	C2	0.93	1.00	0.00	0.00
	C3	0.93	1.00	0.00	0.00
	C4	0.93	1.00	0.00	0.00
	C5	0.92	1.00	0.92	0.84
	C6	0.92	1.00	0.92	0.84
Crane	C1	0.94	1.00	0.38	0.57
	C2	0.94	1.00	0.38	0.57
	C3	0.94	1.00	0.38	0.57
	C4	0.94	1.00	0.38	0.57
	C5	1.00	1.00	0.23	0.09
	C6	1.00	1.00	0.23	0.09
Orange	C1	0.90	0.80	0.00	0.00
	C2	1.00	0.90	0.19	0.00
	C3	1.00	0.90	0.19	0.00
	C4	1.00	0.90	0.19	0.00
	C5	0.90	0.80	0.00	0.00
	C6	1.00	0.90	0.19	0.00
Bass	C1	1.00	1.00	0.45	0.30
	C2	1.00	1.00	0.45	0.30
	C3	1.00	1.00	0.45	0.30
	C4	1.00	1.00	0.45	0.30
	C5	1.00	1.00	0.18	0.20
	C6	1.00	1.00	0.18	0.20

The precision graphs for the above table can be seen as under. Figure 4 and Figure 5 show the precision graph for 'Bank' for page 1 and page 2 respectively, Figure 6 and Figure 7 show the precision graph for 'Crane' for page 1 and page 2 respectively, Figure 8 and Figure 9 show the precision graphs for 'Orange' for page 1 and page 2 respectively, Figure 10 and Figure 11 show the precision graphs for 'Bass' for page 1 and page 2 respectively. All of the precision graphs have been depicted below. These graphs help in evaluating our system's accuracy and gives a pictorial representation of results obtained by the refining the query entered by the user as well as comparing both of their results.

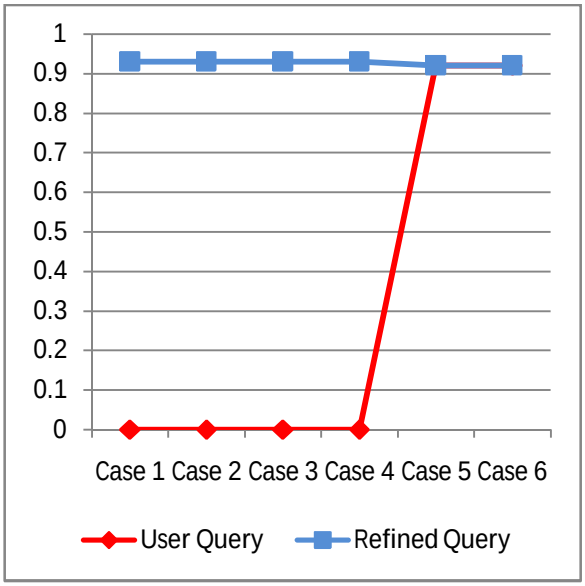

Figure 4: Bank Precision Graph Page 1

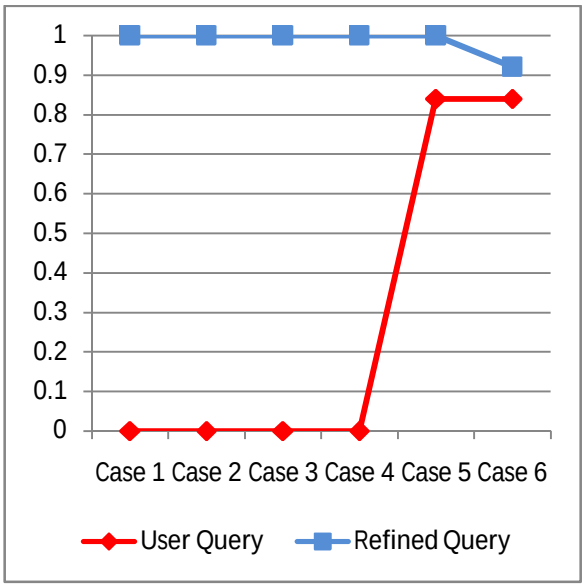

Figure 5: Bank Precision Graph Page 2

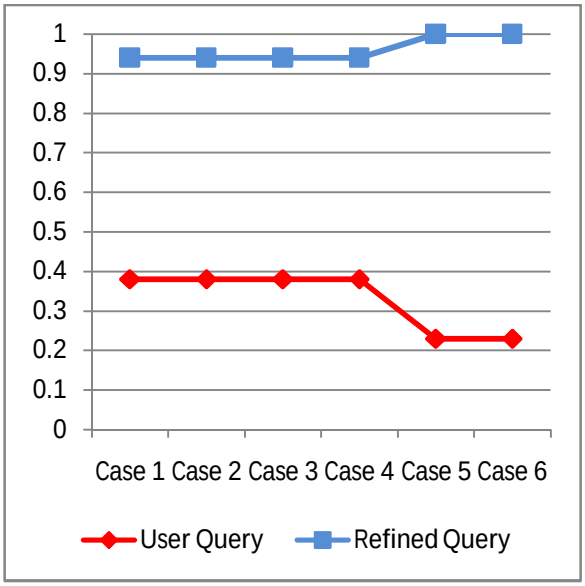

Figure 6: Crane Precision Graph Page 1

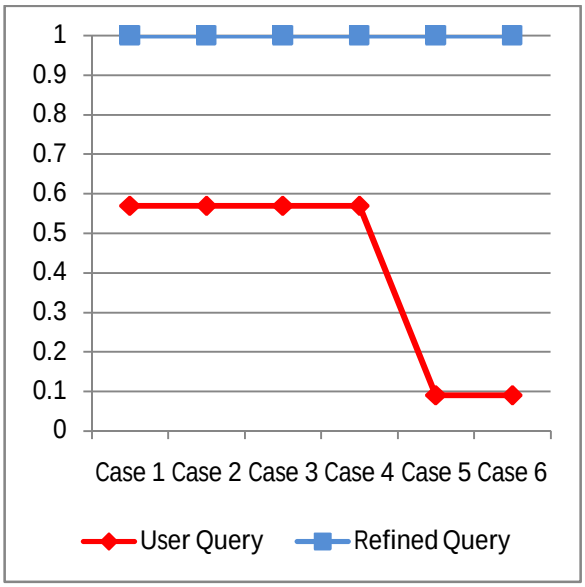

Figure 7: Crane Precision Graph Page 2

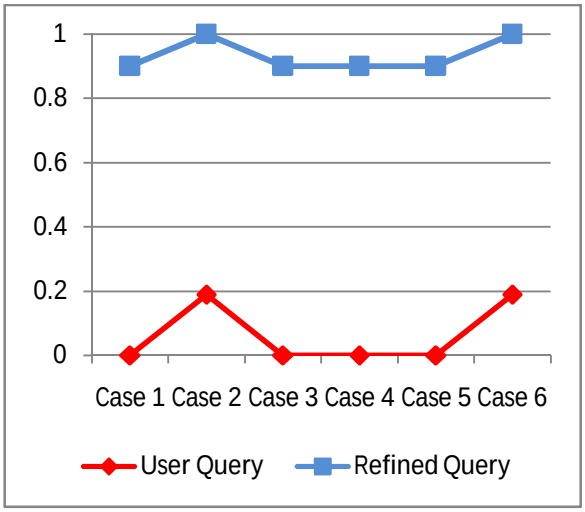

Figure 8: Orange Precision Graph Page 1

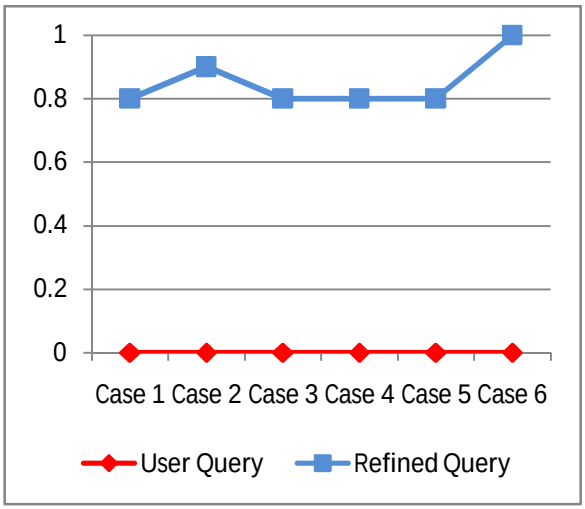

Figure 9: Orange Precision Graph Page 2

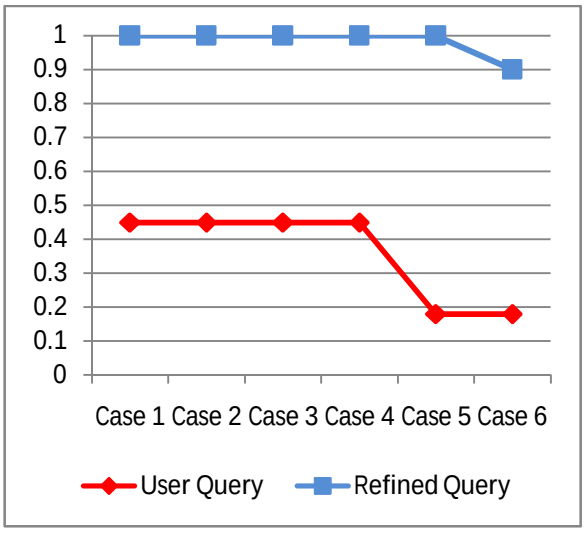

Figure 10: Bass Precision Graph Page 1

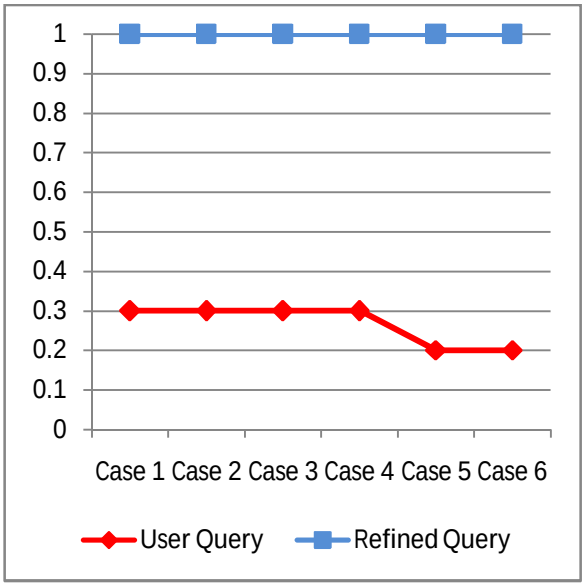

Figure 11: Bass Precision Graph Page 2

9. CONCLUSION

As web is a network of various other networks, information is scattered all over the net. People type keywords for searching content on the web and hence the results are generated based on the keywords entered. An intelligent helps in reformulating the query entered by the user and thus outputs better and refined results. It can be seen in the precision table and graphs depicted above. Hence further improvements and refinements can be performed in this direction so as to gain better results and most importantly save user's time in filtering out relevant results out of the irrelevant ones.

REFERENCES

[1] Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008). **Introduction to Information Retrieval**. Cambridge University Press.

[2] http://www.websearchsg.org/the-history-of-web-search.php

[3] R. Navigli. **Word Sense Disambiguation: A Survey**, ACM Computing Surveys, 41(2), 2009, pp. 1-69.

[4] Eneko Agirre and Aitor Soroa. Semeval-2007 task 02: **Evaluating word sense induction and discrimination systems**. Proceedings of the 4th International Workshop on Semantic Evaluations, p.7-12, June 23–24, 2007, Prague, Czech Republic

[5] http://interscience.in/IJCNS_Vol1Iss4/39-44.pdf