Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.16 – 22

RESEARCH ARTICLE

Study and Implementation of Case and Relation Based Algorithm (CARE) with Pigeon Algorithm

Miss. Neha B. Thakare[1], Prof. R.R.Shelke[2]

[1]M.E. Second Year, CSE, HVPM COET, Amravati University, Amravati, India
[2]Prof. CSE Department, HVPM COET, Amravati University, Amravati, India
[1] nhthakare@gmail.com, [2] rajeshrishelke@rediffmail.com

Abstract - Today's Web is a human-readable Web where information cannot be easily processed by machine. Information retrieval mechanisms from the web become tedious as the amount of content is growing dynamically every day. A New Integrated Case and Relation Based Page Rank Algorithm have been proposed to rank the results of a search system based on a user's topic or query. The semantic web has an idea of connecting, integrating and analyzing data from various data sources and forming a new information flow, hence a web of databases connected with each other and machines. This paper proposes an optimized semantic searching of keywords represent by simulation an ontology with a proposed algorithm which ramifies the effective semantic retrieval of information which is easy to access and time saving by including the novel approach of page ranking by the combination of CARE and Pigeon Algorithms.

Keywords: Case and Relation Based Page Rank Algorithm (CARE), ontology, pigeon algorithm

I. INTRODUCTION

The Web contains a huge amount of data but computers alone cannot understand or make any decisions with this data. Today, search engines constitute the most helpful tools for organizing information and extracting knowledge from the Web [1-5]. The Semantic Web will offer the way for solving this problem at the architecture level. In fact, in the Semantic Web, each page possesses semantic metadata that record additional details concerning the Web page itself [6-8]. Annotations are based on classes of concepts and relations among them.

The "vocabulary" for the annotation is usually expressed by means of an ontology that provides a common understanding of terms within a given domain. In our proposed work we are giving implementing the page ranking concept with more efficient way with the combination of CARE and Pigeon algorithm.

In which the score will be calculated for each of the page to provide the ranking to the pages. The page having maximum value of score will be displayed first and so on. So that the user can understand that which page is more accurate according to their provided keywords.

II. LITERATURE REVIEW

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 indexing [1][2]. Today Search Engines are highly used for the purpose of Information Retrieval and with such a high demand it becomes essential to refine the results and present user with the most appropriate top ranked results. Web Mining is the field which deals with such concepts.

Page rank concept is Proposed by Sergey Brin and Larry Page. They have said that, A web page is important if it is pointed to by other important web pages.

J. Kleinberg [5][6] gave us algorithm "Hyperlink Induced Topic Search" (HITS). He recognized two different forms of web pages called hubs and authorities. Authorities are pages having important contents whereas Hubs are pages that act as resource lists guiding users to authorities. Thus a good hub page for a subject points to many authorities pages on that content, a good authority page is pointed by any good hub pages on the same subject.

Kleinberg [20-22] said that a page may be a good hub and good authority at the same time and this circular relationship lead to the definition of an iterative algorithm HITS [5-16]. This algorithm was used in search engine "CLEVER" but was not successful because of topic drift and efficiency problems. Also this algorithm worked on both Web Structure Mining as well as Web Content Mining.

In case of CARE algorithm, Relation based page ranking algorithm also has the same role in search engine but it is the graph based approach and hence used to require more time than the CARE. It requires more execution time than CARE. This can be graphically shown as follows.

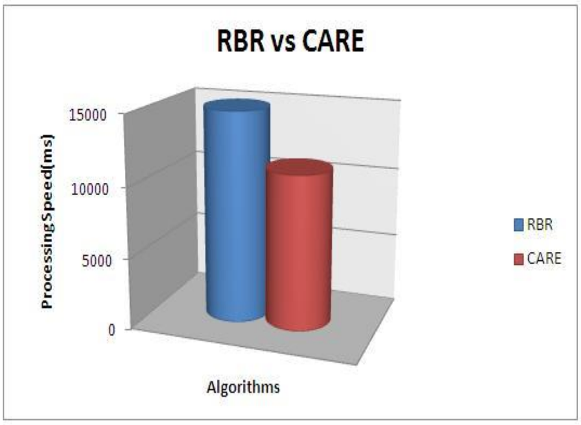

Fig. 1.Time consumption of RBR and CARE

The unavoidable drawback of CARE algorithm is that, the accuracy and the time complexity cannot be maintain with the CARE algorithm alone. It must be accomplished with any optimization technique to increase its efficiency i.e. nothing but the pigeon page rank algorithm.

In February 2011, the first search filter that was part of the Panda update was rolled out. It's basically a content quality filter that was targeted at poor quality and thin sites in order to prevent them from ranking well in Google's top search engine results pages (SERPs).

Panda 1.0 update: The search filter was aimed at content farms — those sites that employ many writers who create poor quality content around specific keywords in a bid to rank in Google's top ten results. This update was primarily aimed at U.S. sites and affected 12% of search results.

III. PROPOSED WORK

In our proposed work, we have developed the Case and relation based algorithm by adding the subpart of pigeon algorithm.

A. CARE Algorithm

The CARE algorithm is a new algorithm proposed to constitute the power of relationships and previous history. Relation based page ranking algorithm produce more accurate results. Thus various search engines are presented for better information extraction by using relations of the semantic web. In this approach it generates annotated page graphs for all the resulting pages generated by the search engine and then it ranks those pages if the number of result pages is very large then it takes more time to generate the annotated page graphs for each result page.

The CARE algorithm is using Textual Case Based Reasoning (TCBR) [7] and Relation-based Page Ranking algorithms. The textual case based reasoning is used to reduce the number of incorrect result pages during the search process. Textual case based reasoning uses previous knowledge of the search results by using this it fetches the result for the search query which are more relevant to the search query then after this we generate the annotated page graphs for these result pages[16-20]. By this the search time can be reduce by avoiding the generation of annotated page graphs for irrelevant results.

B. PROBLEM ANALYSIS

Many search engines are having major drawback as it lacks interpretability between machines, metadata and knowledge management crisis. Powerful and complex algorithms are required by the search engines in order to parse the keywords requested by the user. The future web, semantic web is based on the principal of interoperability between machines and giving them power to think, aims at attaching metadata, specifying relations between web resources and knowledge management, in order to process and integrate data by the users[2].

For example, Let a user types a keyword set mountain, Then by looking at the texts found by means of mountain, some other words related to mountain can be determined like antonyms and synonyms. Obviously, these texts may not include any query word and they include merely one related word or more[1]. This aims at finding the probabilities of the keywords with the maximum likelihood occurrence if keywords in a set.

IV. PROPOSED METHOD

The proposed work is based on PageRanking concept. For this, we apply Case and relation based algorithm accompanied by the subpart of the pigeon algorithm.

A. Study of various search data-sets

In proposed design, first we study the various search datasets, which can be used to evaluate the project work, i.e. the datasets must include text and numerical data for evaluation of Page Rank algorithm. Then we will study and implement of CARE Algorithm**.**

B. Study and implementation of CARE Algorithm

In this module, we would be developing the CARE algorithm which would help us to search the dataset collected in module 1, and based on the input search query, re-rank the results present in the dataset, and produce the re-ranked output. The delay and accuracy of the care algorithm on dataset would be studied

C. Optimization of the CARE algorithm by adding a sub part of Google's Pigeon Rank System **:**

The Google's Pigeon rank system would be studied and probable optimizations to the CARE algorithm would be find out and Pigeon rank algorithm applied so that we could get better, more optimized and accurate outputs.

D. Comparison of algorithms and result evaluation **:**

In this module, the CARE and the optimized CARE algorithms would be compared and the results would be checked and optimized if required. The following figure exhibits the comparison between Relation based Page Ranking and Case and Relation based Page Ranking. There are two types of evaluation made to confirm the effectiveness in terms of execution time and the accuracy of the search result.

Our proposed method consists of the combination of Case and relation based algorithm (CARE) and pigeon algorithm.

V. CARE ALGORITHM

The CARE algorithm is a new algorithm proposed to constitute the power of relationships and previous history. Relation based page ranking algorithm produce more accurate results. Thus various search engines are presented for better information extraction by using relations of the semantic web. In this approach it generates annotated page graphs for all the resulting pages generated by the search engine and then it ranks those pages if the number of result pages is very large then it takes more time to generate the annotated page graphs for each result page.

- Implementation of the CARE Algorithm

```
Input : User Search Query(Q)
Output : Set of Web pages satisfy User Query
Procedure : CARE Algorithm
CARE (Q)
Begin
ho _ set of hyponym (user query)
he _ set of hypernym (stored in knowledge base)
hse _ set of hyper-hyponym relation
hspe _ set of pruned hyper-hyponym relation
G(C,R) _ Ontology graph
where C − Concepts (nodes) in the Ontology graph and
R − Edges (Relations) between the nodes in the Ontology graph
Gq (Cq , Rq)_ Query sub-graph
where Cq − Concepts (nodes) in the Query sub-graph and
Rq − Edges (Relations) between the nodes in the Query sub-graph
Ga (Ca ,Ra) _ Annotated graph
where Ca − Concepts (nodes) in the annotated page graph and
Ra − Edges (Relations) between the nodes in the annotated page graph.
Gp(Cp,Rp) _ Page sub-graph
where Cp − Concepts (nodes) in the Page sub-graph and
Rp − Edges (Relations) between the nodes in the Page sub-graph. Foreach Page
sub-graph do
Begin
Label the edged in Gp with an index ranging from 1 to Rp
Define variable e and a to index graph edges
Set _e = _ij
Set _e = _ij
Set _e = _ij / _ij
Mark all the edges in Gp as not visited
Allocate weight vector W of size |Cp|-1
Allocate vector _ of size |Cp|-1
Initialize W and _ to zero
for e=1, e_|Rp|, e=e+1
Begin
mark edge e as
visited visit (e,e,1,_e)
W[1] = W[1] + _e
_[1] = _[1] + 1
                    End
               End
          End
Visit (o,e,l,s)
Begin
a = e + 1
while a_|Rp| and l_|Cp|-1
Begin
If a is not visited and a is safe then
Begin
mark edge a as visited visit (o,a,l+1,s×_) W[l+1]
= W[l+1] + s
_[l+1] = _[l+1] + 1
set edge a as not visited
                End
Else
a = a+1;
          End
    End
```

- The CBR Cycle

At the highest level of generality, a general CBR cycle may be described by the following four processes [10].

A new problem is solved by retrieving one or more previously experienced cases, reusing the case in one way or another, revising the solution based on reusing a previous case, and retaining the new experience by incorporating it into the existing knowledge-base (case-base) [14]. The four processes each involve a number of more specific steps, which will be described in the task model. In figure 3, this cycle is illustrated.

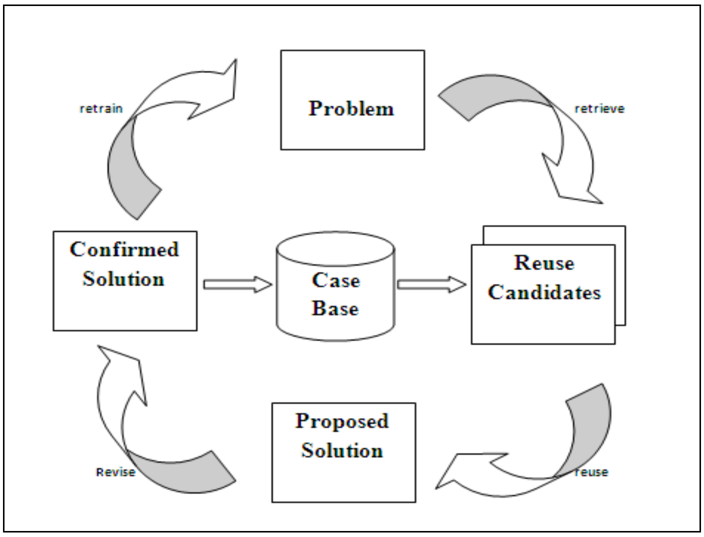

Fig. 2. The CBR Cycle

VI.PIGEON ALGORITHM

Pigeon algorithm is a new algorithm to provide more useful, relevant and accurate local search results that are tied more closely to traditional web search ranking signals. Google stated that this new algorithm improves their distance and location ranking parameters. This algorithm is use for the optimization purpose. This algorithm demonstrates its feasibility to find the shortest path from a given source [3]. This algorithm is based on the intelligence of passenger pigeons and to utilize it for finding shortest path from a given source. It involves Dijkstra's [3] Algorithm for finding the shortest path. Pigeon Rank's success relies primarily on the superior trainability of the domestic pigeon and its unique capacity to recognize objects regardless of spatial orientation. The common gray pigeon can easily distinguish among items displaying only the minute differences, an ability that enables it to select relevant web sites from among thousands of similar pages.

By collecting flocks of pigeons in dense clusters, Google is able to process search queries at speeds superior to traditional search engines, which typically rely on birds of prey, brooding hens or slow-moving waterfowl to do their relevance rankings.

When a search query is submitted to Google, it is routed to a data coop where monitors flash result pages at blazing speeds. When a relevant result is observed by one of the pigeons in the cluster, it strikes a rubber-coated steel bar with its beak, which assigns the page a Pigeon Rank value of one. For each peck, the Pigeon Rank increases. Those pages receiving the most pecks, are returned at the top of the user's results page with the other results displayed in pecking order.

A. Algorithm

Web crawling:

- Use depth-first or breadth-first search to:

Learn the structure (vertices+ edges) of the graph

- Build an index of the web:

Use a hash table to store pairs (word, list-of-sites)

for each web site S do

```
for each word w in S do
index.get(w).addLast(S)
```

VII. DEVELOPMENT OF FORMULA TO RANK THE PAGES

In our proposed work, we are using the following formula for calculating the scores for the pages on the basis of which the ranks are to be provided to the pages.

$$\text{Score} = [N / L * 100] + C \quad \ldots. (1)$$

Here,

N -No. of times the document test is matching.

L - Document length.

C - Hits for that document.

In our proposed work the score is calculated by considering the all three factors such as N, L, As we are taking the notice of three different factors that's why the value of score will not depend only on any factor like literature work. From the calculated values of score, the page having maximum value amongst all will be displayed first.

Consider the following example, If the page is having 2 lines i.e. the length of document is 2.

L = 2 , N = 1, C = 1, Putting these values in equation 1,we get

Score = [1/2 * 100] + 1

Score = [50] + 1

Score = 51

VIII. RESULT AND DISCUSSION

The proposed technique generates the efficiently ranked pages as very accurate output. Each page is ranked according to its score. The page having maximum number of score amongst all is listed first which shows that it is the most accurate data provided against the query of the user. As three components are involved while ranking the pages , so ranking doesn't depend only on one factor most that is why the produced result is more accurate.

If we compare the literature work with our proposed work then we can conclude that the our result is more accurate than the previous work. This can be graphically shown as follows:

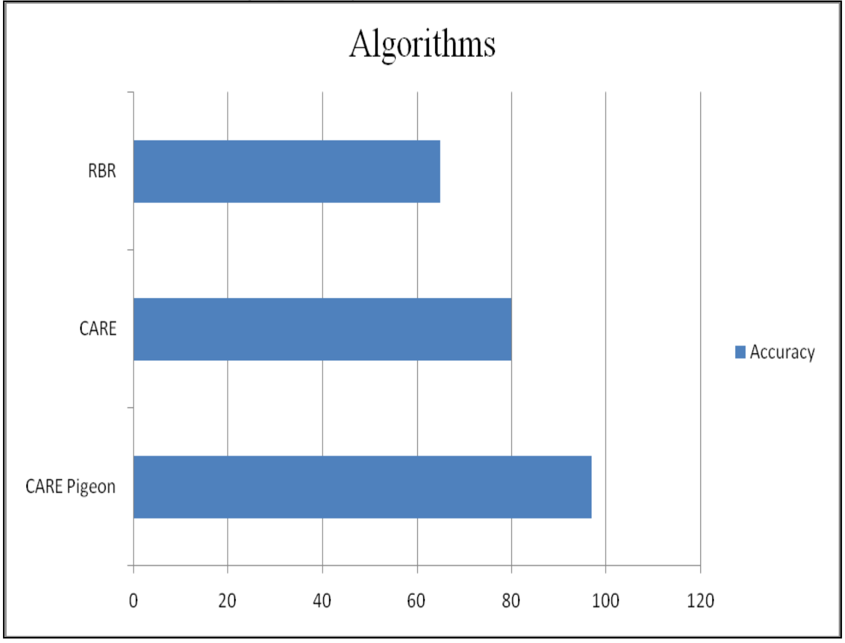

Fig 3: Accuracy between RBR, CARE and CARE + Pigeon Algorithms

In the above graph plots the accuracy comparison for the three algorithms. The accuracy for the CARE algorithm alone is 80% while the accuracy for the same algorithm[1] by adding the subpart of the Pigeon algorithm is 97%. So from this we can say that, the implementation of our proposed work produces more accurate outputs than the literature

IX. CONCLUSION

From our proposed work we can conclude that the ranking approach by the combination of CARE and Pigeon algorithm gives more relevant results to the user. Also the ranking is done through the value of score and no. of hits too with respect our proposed work.From above experimental results it concludes following points:

- In the proposed page ranking, we are using Case and Relation Based algorithm and Pigeon algorithm in combination to increase the efficiency of the output.
- From the experimental results, we can conclude that our approach leads to the increased accuracy of outputs.
- Leads to better performance than the PageRank, HITS, CARE, RBR algorithms[21-23].

X. FUTURE SCOPE

In the future, the results of proposed PageRank approach can be studied under different metrics with other PageRank algorithms. Further the project can be modified by choosing more and more accurate formula for evaluating scores of the pages and also the time required for the ranking can also be reduce.

ACKNOWLEDGEMENT

I am Miss Neha B. Thakare graduated in 2012 with a BE degree in Computer Science and Engineering and now appeared to the Second Year of Master degree in H.V.P.M. C.O.E.T. Amravati.

Prof. Rajeshri R. Shelke is Professor of Department of Computer Science and Engineering in H.V.P.M. C.O.E.T. Amravati.

REFERENCES

[1]. A guide to Future of XML, Web Services and Knowledge Management by Michael C.Daconta, Leo J. Obrst, Kevin T.Smith,2003

[2]. Ms.N.Preethi, Dr.T.Devi, New Integrated Case And Relation Based (CARE) Page Rank Algorithm, ICCCI -2013

[3]. Hang Sun, and Haibin Duan, *Senior Member,* PID Controller Design Based on Prey-Predator Pigeon-Inspired Optimization Algorithm, IEEE 2014.

[4]. LI yuan, ZENG jianqiu, Web 3.0: A real personal web! Beijing China, IEEE 2009

[5]. Radha Guha," Towards the Intelligent Web Systems, Coimbator (India)," IEEE 2009

[6]. J. Kleinberg,"Authoritative Sources in a Hyper-Linked Environment", Journal of the ACM 46(5), pp. 604-632, 1999. [7]. J. Kleinberg, "Hubs, Authorities and Communities", ACM Computing Surveys, 31(4), 1999.

[8]. Toby Segaran, Colin Evans, Jamie Taylor,"Programming the Semantic Web by, O‟REILLY",IEEE 2009

[9]. Dean Allemang, "Rule-based intelligence in the Semantic Web," TopQuadrant Inc. IEEE 2006

[10]. A.Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and systemapproach," AI Communicatons , vol. 7, no.1, pp. 39–59, 1994

[11]. J.L.Kolodner, Case-Based Reasoning Morgan Kaufmann: San Mateo, CA,1993, pp. 27–28.

[12]. I. Watson, "Applying Case-Based Reasoning: Techniques for Enterprise Systems," Morgan Kaufmann San Francisco,CA,1997.

[13]. Althoff, K.D (1989). Knowledge acquisition in the domain of CNC machine centers; the MOLTKE approach.In John Boose, Brian Gaines, Jean- Gabriel Ganascia (eds.): EKAW-89; Third European Workshop onK nowledge-Based Systems, Paris, July 1989. pp 180-195.

[14]. J.Kolodner, "Making the implicit explicit: Clarifying the principles of case-based reasoning", Experiences, Lessons & Future pp. 349-370, AAAI Press, Menlo Park, USA, 1996.

[15]. W.Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proc. Of the Second Annual Conference on Communication Networks and Services Research (CNSR ‟0), IEEE, 2004.

[16]. Fabrizio Lamberti, Andrea Sanna and Claudio Demartini, "A Relation-Based Page Rank Algorithm for Semantic Web Search Engines", In IEEE Transaction of KDE, Vol. 21, No.1, Jan 2090.

[17]. T.Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web,"Scientific Am., 2001

[18]. J.L. Kolodner, Case-Based Reasoning, Morgan Kaufmann: San Mateo, CA, 1993, pp. 27–28.

[19]. RekhaJain, Rupal Bhargava and G.N Purohit, Ambiguity Resolution in Information Retrieval D.Gentner: Structure mapping - a theoretical framework for analogy. Cognitive Science, Vol.7. s.155-170. 1983

[20]. B.H Ross: Some psychological results on case- based reasoning. Reasoning Workshop , DARPA 1989.

[21]. R.Schank: Dynamic memory; a theory of reminding and learning in computers and people. Cambridge University Press. 1982.

[22]. J.L.Kolodner, Case-Based Reasoning, Morgan Kaufmann: San Mateo, CA, 1993, pp. 27–28

[23]. Janet L. Kolodner, An Introduction to Case Based Reasoning, College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0280,U.S.A