UGC JOURNAL NUMBER 44557

International Journal of Allied Practice, Research and Review

Website: www.ijaprr.com (ISSN 2350-1294)

Semantic Web Mining using Shannon Information Gain

Doshi Poonam Pradhumnakumar and Dr. Emmanuel M.
Research Scholar, Pacific University, Udaipur, Rajasthan, India
PICT, Pune, Maharashtra, India

Abstract- Hundreds of millions of information are generated each day and as such billions of web pages are created coagulating this enormous data with various web links connected to them. In order to answer millions of queries relating to these colossal numbers of web pages, we need to optimize the performance of the Search Engine, which is assigned with the task of fetching the relevant information pertaining to the queries asked.

The proposed research experiments have been conducted on Shannon information gain to determine the threshold value of dynamic dataset. Cosine similarity and Shannon information gain ratio are two important factors to achieve the result on the seed URL.

Keywords: ***Search engine, crawling, Shannon information gain.***

I Introduction

In Current Scenario Web is source of vital information. Almost 4.05 billion pages been indexed by search engines and total count of indexed pages remains non- estimated from 2006. Searching and Retrieving relevant information is major objective of computer technology. Finding new patterns from web and retrieving information related to topic are major challenges for web mining domain. Commercial search engines like Google, Bing have proved to be the best ones for Information search on web and process almost 2.5 petabytes of data daily.

Outline and Development of better Web creeping framework for finding new inventive examples from web is investigate challenge. A superior and Effective recovery of data and finding fascinating examples from web, web crawlers are required. Crawlers support in retrieving information related to seed informationgiven by user and make process of indexing better.Crawler is automatic software program which help in indexing of web pages. According to the technique involved in crawling operation, the crawler systems are classified into different types. A search engine has to navigate through a vast number of pages in order to build and maintain a useful list of words.The paper prime objective is of applying cosine similarity to get the exact match of the similar words, Shannon information gain ratio to increase the performance and its scalability. Due to its emphasis on the usage of a predominantly centralized crawler, Traditional crawling methods suffer from a severe

limitation on its reconfiguration capability. This has given rise to the evolving study of crawling by employing the semantic data.

II Literature Review

The crawler based on semantic data extraction has the capability to traverse the gigantic web, down loading the relevant data. Industrial automation has wide exposure to Semantic technologies. Self-adaptive semantic focused crawler – also named the SASF crawler [1], is dedicated to the purpose of identifying, refining and indexing mining service precisely and efficiently over the Internet. A technology of semantic based crawling is encompassed into the semantic framework in order to maintain the performance of crawler, regardless of the variety in the Web environment.The crawling operation is performed through the process called URL ordering [2]. This is useful in crawling the web. However, it needs to revisit the pages in order to ascertain the changes made thereafter.

Arotaritei, Dragos[3], have written different techniques used for the purpose of web mining are highlighted. The authors have contended that a computing tool would be helpful in reaching to precise and relevant information from the vast pool of unstructured log files accumulated during the web mining process. Number of web mining systems such as semantic search, text and image retrieval, clustering, recommendation and many more are better elaborated [4].

To provide an overview of how to use frequent pattern mining techniques for discovering different types of patterns in a Web log databases [5]. The goal of discovering frequent pattern in web log data is to obtain information about the navigational behavior of the user. It is very difficult to deal with this behavior of the user, so a better system is implemented to give error free output.

We have studied research work as given by several researchers and have come out with the conceptual description of the different strategies of web crawling where different papers with different crawling methods are implemented and it is given as below:

Table 1: Different types of web crawler strategies [6]

References Title	**Methods**	**Concept**	**Advantage**
1. Scheduling algorithms for Web crawling [7].	Crawling the large sites first.	Crawling starts with the sites with the large number webpages which have not been traversed yet.	Large web sites crawled first.
2. Effective Page Refresh Policies for Web Crawlers_ ACM Transactions on Database Systems [8].	Breadth First Search Algorithm.	Starts at the root URL and searches the all the neighbors URL at the same level.	Well suited for situations where the objective is found on the shallower parts in a deeper tree.
3. Artificial Intelligence illuminated [9].	Depth First Search Algorithm.	Starts at the root URL and traverse deep through the child URL.	Well suited for problems.

4. Graph theory with applications to engineering and computer science [10].	Page Rank Algorithm.	Download the web pages on the basis of page-rank.	In the very limited time important pages are downloaded.
5. Semantic Similarity Based Focused Crawling [11].	Online Page Importance Calculation Algorithm.	The crawler will download web pages with higher cache memory in each stage and this memory range will be distributed between the pages it points when a page is downloaded.	The cash value is calculated in one step and in short duration of time.
6. A Query based Approach to Reduce the WebCrawler Traffic using HTTP Get Request and Dynamic Web Page [12].	By HTTP Get Request and Dynamic Web Page.	It is a query based approach in which crawler just download dated webpages after the last visit.	Web crawler download only downloads latest updated webpages.
7. A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency [13].	By the use of Filter.	In this query based approach, crawling is done by the use of filter.	Reduces network traffic.
8. Efficient crawling through URL ordering [14].	Crawling through URL Ordering	It visits earlier URLs that have anchor text which is similar to the driving query or link distance is also short to a page.	Extremely useful when we are trying to crawl a fraction of the Web, and we need to revisit pages often to detect changes.

III. Proposed System

In the initial stage proposed system accepts the seed URL from the user. System spiders the URL for information and stores it for the future uses. There are many of applications of web pages but in general their working is divided into three steps as follows: download the web page, parse the web page for content and finally, parse the outgoing sub links. The architecture of the system is as shown in Figure 1:

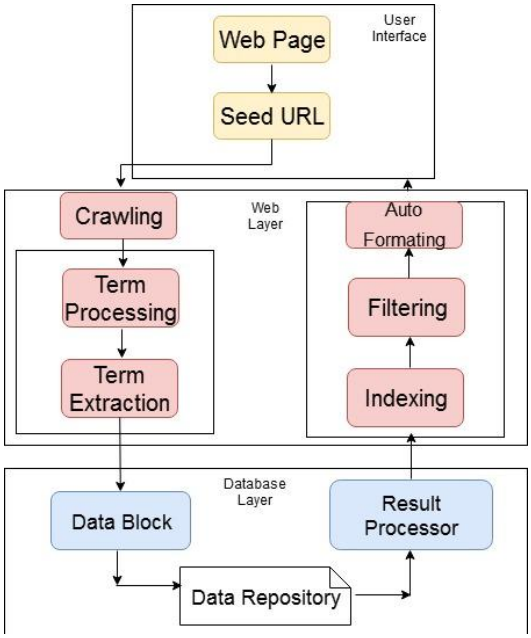

Figure 1: Proposed Architecture system

Semantic focused crawling technology is utilized to resolve the matters of mining service information.

Let S={ } be a system for pattern recognition for web information.
A set S = {s1, s2, s3,….sn} where sn is seed URL.
To identify the interesting pattern S={sn, I}.

The system is divided into different stages and implemented with the help of algorithm to produce the pattern matching.Further Depth-first algorithms are viewed as two measurements to make decisions on what URLs to decide for the next indexing process. There are semantic similarity metrics based on previous works including cosine similarity and edit distance. These measurement have the following definitions: m1 and m2 are two vectors of attributes, the cosine similarity, cos(m1,m2) represent using a dot product and magnitude using the following formulae:

m1.m2=||m1||||m2|| cos(m1,m2) ……...……(1)

$$\cos(m1, m2) = \frac{m1.m2}{||m1||||m2||} \quad \text{……...……(2)}$$

Proposed crawler employs the DFS algorithm to gather information.By travelling through the search and by starting at base web page and then navigating through child web links in much deeper level, the Depth First Search (DFS) serves as a useful technique. It starts from the leftmost child whois fetched first and then the right hand side to traverse each node. This process keeps on visiting the nodes till it cover last node. To include, the crawler should first visit the recently added from the request queue. It then travels through the node which has not been visited yet. This process is continued until all the nodes have been visited. This particular algorithm is more appropriate to search problems. However it enters the infinite loop in the wake of large no of branches.

Algorithm: Depth First Crawlers

graph G and a start vertex node of G
SetLabel(node, VISITED)
edge2 G.incident Edges(node)
GetLabel(edge) = UNEXPLORED
weight = opposite(node; edge)

GetLabel(weight) = UNEXPLORED
SetLabel(edge, DISCOVERY)
DFS(G, weight)
SetLabel(edge, BACK)

Output: The edges of G in the connected component of node are labeled as discovery edges and back edges.

Shannon Information Gain: Computers have simplified the concept of bit, a unit of information that takes two values, 0 or 1. Shannon information uses the same. The input to Shannon Information Gain is given in the form of blocks. The value of Shannon Information Gain is in between 0 and 1. The formula to calculate Shannon Information Gain:

IG(O) = (F / S) log (F / S) - (P / S) log (P / S) …….(3)

Where F = Frequency of the present count
P = Non presence count
S = Cluster Elements Size.
IG(O) = Information Gain for the given Object

For every word support value i1s calculated and displayed for searching faster data. Bubble sort algorithm is applied on it. Sorting is done on support value and not on word.

IV. Results and Discussion

The system is Implementation in java on windows platform and test for dynamic web pages. The live data is taken from seed URL crawled which is given for pattern matching items by using cosine similarity and Shannon information gain. By using the library support of Jaya we have coded in Java platform. System configurations: Operating System: Windows XP, RAM size: 2GB, Processor: Intel Core 2 Duo and Coding Platform: Java 1.6, Eclipse.The Original dynamic crawled data is further given for finding out the important keywords, nothing but for pre-processing of data. The help of pre-processing algorithm all the unwanted words are illuminated.

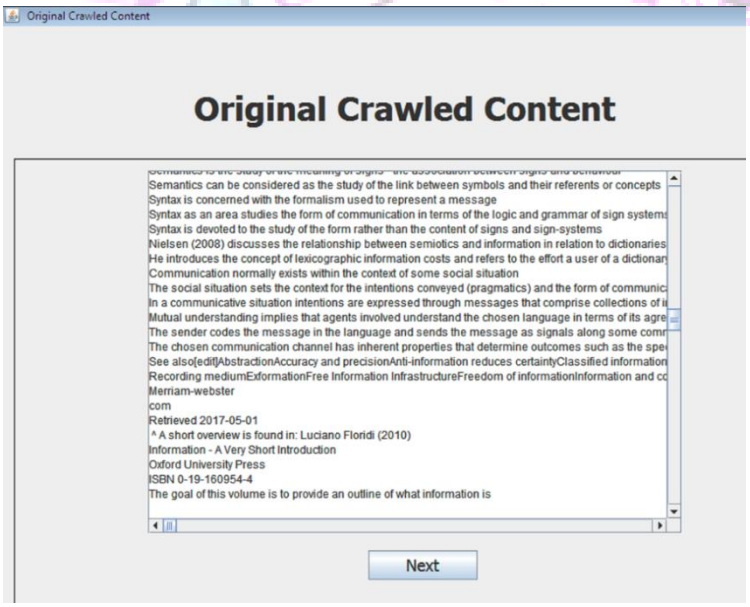

Figure 1: The original crawled data

Cosine similarity is one of the two metrics utilized by the Depth First Search (DFS) algorithms. The cosine similarity in the two words “magnify” and “magnification” is calculated to find the similarity metrics which are shown in tabular form is as given below in figure 2.

Name	Type	Value
A[ ]	Array list	Size = 7
A[0]	Character	M
A[1]	Character	A
A[2]	Character	G
A[3]	Character	N
A[4]	Character	I
A[5]	Character	F
A[6]	Character	Y

Figure 2: Word array for magnify

In the first step, first word 'magnify' is placed in an array with each cell carrying one letters each. The two words "magnify" and "magnification" is compared with it and then the union of the two words is calculated. The resultant word emerged in the process have the following characters: MAGNIFYCATION which is shown in the figure 3.

Name	Type	Value
A[ ]	Array list	Size =13
A[0]	Character	M
A[1]	Character	A
A[2]	Character	G
A[3]	Character	N
A[4]	Character	I
A[5]	Character	F
A[6]	Character	Y
A[7]	Character	C
A[8]	Character	A
A[9]	Character	T
A[10]	Character	I
A[11]	Character	O
A[12]	Character	N

Figure 3: Word array for 'magnifycation'

The binary vectors are produced in view of the union of the two words. We assign 1 to the vector cell if the word and the union have the same character and if they don't, we assign 0. For example, using the word "magnify," based, we see that the m,a,g,n,i,f, are all 1 because both arrays have the same values. However, characters of y,c,a,t,i,o,n are 0 because the magnify does not have these characters. The resultant vector attained in the process is generated based on the union and the word "magnify".The cosine similarity is calculated by multiplying each cell with the corresponding cell shown in table 1:

Table 1: 1-Words Vectors

VW1	1	1	1	1	1	1	1	0	0	0	0	0	0
VW2	1	1	1	1	1	1	0	1	1	1	1	1	1
Dot	1	1	1	1	1	1	0	0	0	0	0	0	0

$\overrightarrow{VW1} \cdot \overrightarrow{VW2} = 6$ ..... (4)

Then we calculate the cosine similarity with the given formula:

$$\frac{\overrightarrow{VW1}.\overrightarrow{VW2}}{||\overrightarrow{VW1}||\,||\overrightarrow{VW2}||} = \frac{6}{||1.4284||\,||6||} = 0.70011668611435 \quad \ldots\ldots (5)$$

Furthermore, portion of the results that were exported in the Excel spread sheet is shown in the following table,

Table 2: Portion of the URL's

Sr. No	Parent URL	URL	Keywords	Distance
1	https://en.wikipedia.org/wiki/Social_science	https://en.wikipedia.org/wiki/Social_studies	Education, humanities,	0.560070812
2	https://en.wikipedia.org/wiki/Social_science	https://en.wikipedia.org/wiki/Main_Page	eruption, geology, volcanoes	0.587212511
3	https://en.wikipedia.org/wiki/Social_science	https://en.wikipedia.org/wiki/Portal:Featured_content	feature, collaborative	0.332057812
4	https://en.wikipedia.org/wiki/Social_science	https://en.wikipedia.org/wiki/Portal:Contents	Biology, Human Genome	0.467652744
5	https://en.wikipedia.org/wiki/Social_science	https://en.wikipedia.org/wiki/Scientific_socialism	Scientific, political_economic_	0.445216785

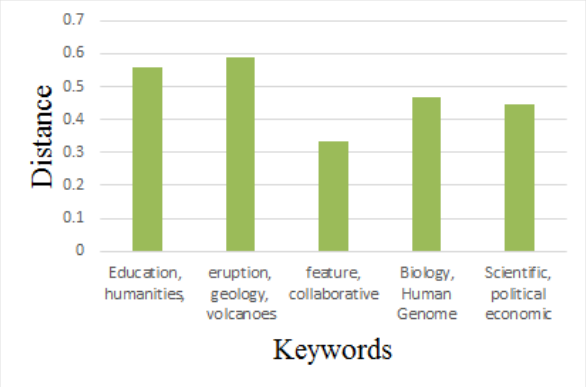

Figure 4: Distance calculated for URL keywords

This value is arranged in a sorted manner after the calculation of the gain ratio. For every word support value is calculated and displayed for searching faster data. Bubble sort is implemented to sort the data in ascending order. The following table shows the Shannon information gain.

Table 3: Shannon information gain ratio on dynamic dataset

Important words	**Gain Ratio**
stance lifestyle meplex mental	0.027
link seditlook	0.027
communication organizational	0.049
Analysis	0.086
Freeman	0.058
Information	0.663
External	0.049
Representation	0.162
University	0.068
Represents	0.055
Effective	0.035
Encyclopedia	0.178
Quantitative	0.169
anthropology	0.19

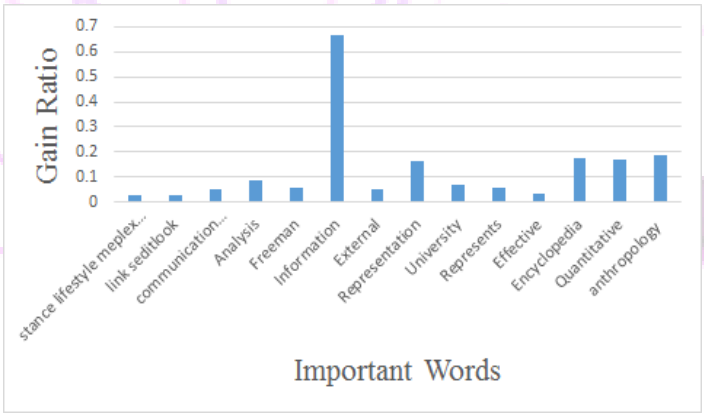

Figure 5: Analysis of Shannon information gain ratio on dynamic dataset

Figure 5 explains the analysis of important words and its Gain ratio, where we get this for the different words which are pre-processed after crawling the data. The gain ratio for information word in the figure achieves the highest rage as 0.663. The data achieved in very less time of period. In this way the proposed system shows the increase in performance andits scalability.

V. Conclusion

The research work presented in this paper was motivated by day to day increase in the internet data. It was also motivated by understanding the traditional and structured crawler. The proposed research work is done on optimizing the performance of the Search Engine. Research experiments on dynamic crawled dataset by calculating the cosine similarity and Shannon information gain threshold value. Cosine similarity is calculated on the similarity value in between similar words, as we have seen for two words “magnify” and “magnification”, the similarity value is 0.7001167 Shannon Gain algorithm is then applied to demystify the threshold values which showcase the frequency of the semantic term being utilized in the shortlisted pages. The comparative analysis of

various word pattern applied to this structured data gives us holistic picture. This analysis which is performed with the threshold time set ranging between 0 and 2.This work could be extended to the other domains by conjoining this work with the optimized interface.

VI. References

[1] H. Dong and F. K. Hussain, "Self-adaptive semantic focused crawler for mining services information discovery", *IEEE Transaction on Industrial Informatics,* vol.10, no.2, pp.1616–1626, 2014.

[2] J. Cho, "Efficient crawling through URL ordering", *Computer Networks ISDN System*, vol.30, no.1–7, pp.161–172, 1998.

[3] D. Arotaritei and S. Mitra, "Web mining: A survey in the fuzzy framework", *Fuzzy Sets System.(Elsevier)*, vol.148, no.1, pp.5–19, 2004.

[4] E. Hüllermeier, "Fuzzy methods in machine learning and data mining: Status and prospects", *Fuzzy Sets System (Elsevier)*, vol.156, no.3, pp. 387–406, 2005.

[5]T. Slimani and A. Lazzez, "Efficient Analysis of Frequent itemset Association Rule Mining Methods", *International Journal of Scientific & Engineering Research,*vol.6, no.4, 2015.

[6]Poonam P. Doshi, Emmanuel M, "Web Pattern Mining Using Eclat", International Journal of Computer Applications (0975 – 8887), New York, Volume 179 – No.8, pp.9-14, Dec2017.

[7] C. Castillo, M. Marin, and A. Rodriguez, "Scheduling algorithms for web crawling", *WebMedia LA-Web Joint Conference 10th Brazilian Symposium on Multimedia and the Web 2nd Latin American Web Congress, Proc.*, pp.10–17, 2004.

[8] J. Cho and H. Garcia-Molina, "Effective page refresh policies for Web crawlers", *ACM Transaction on Database System*, vol.28, no.4, pp.390–426, 2003.

[9] B. Coppin, "Artificial Intelligence Illuminated", *JONES AND BARTLETT PUBLISHERS*, pp.1-768, 2004.

[10] N. Deo, "Graph Theory with Applications to Engineering and Computer Science Paperback – 1979", Prentice Hall India Learning Private Limited, 1979.

[11] M. Ravakhah and M. Kamyar, "Semantic similarity based focused crawling", *2009 1st International Conference on Computational Intelligence, Communication Systems and Networks, CICSYN 2009*, pp.448–453, 2009.

[12] S. Mishra, A. Jain, and A.K. Sachan, "A Query based Approach to Reduce the Web Crawler Traffic using HTTP Get Request and Dynamic Web Page", *Inernational Journal of Computer Application*, vol.14, no.3, pp.8–14, 2011.

[13] S. S. Vishwakarma, A. Jain, and A. K. Sachan, "A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency", *International Journal of Computer Applications (0975 – 8887)*, vol.46, no.1, pp.34–37, 2012.

[14] J. Cho, "Efficient crawling through URL ordering", *Computer Networks ISDN System*, vol.30, no.1–7, pp.161–172, 1998.

[15] M. Emmanuel, S. M. Khatri, and D. R. R. Babu, "A Novel Scheme for Term Weighting in Text Categorization: Positive Impact Factor", *IEEE International Conference System Man and Cybernetics*, pp.2292–2297, 2013.

[16] Y. K. Woon, W. K. Ng, and E. P. Lim, "A support-ordered trie for fast frequent itemset discovery", *IEEE Transaction on Knowledge & Data Engnieering.*,vol.16, no.7, pp.875–879, 2004.

[17] X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance", *Pattern Analysis &. Application*, vol.13, no.1, pp.113–129, 2010.

[18] M. Song, S. Rajasekaran, and S. Member, "For Frequent Itemsets Mining", IEEE, *Knowledge Creation Diffusion Utilization.*, vol.18, no.4, pp.472–481, 2006.

[19] T. A. Kumbhare and S. V. Chobe, "An Overview of Association Rule Mining Algorithms", *International Journal of Computer Science and Information Technologies*, vol.5, no.1, pp.927–930, 2014.