IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Semantic Matching Concept Using Semi-Automated Semantic Algorithm

D.Elangovan[1*], K.Nirmala[2]

[1,2]Manonmanium Sundararanar University, Tirunelveli, India

Corresponding Author: delangomca2000@gmail.com Tel.: 9003350529

DOI: https://doi.org/10.26438/ijcse/v7i2.203206 | Available online at: www.ijcseonline.org

Accepted: 21/Feb/2019, Published: 28/Feb/2019

Abstract: - Semantic matching is a kind of ontology matching technique that depends on linguistics info encoded in light weight ontologies to establish nodes that square measure semantically connected. Ontologies matching are associate operator that identifies those nodes within the two structures that semantically correspond to at least one another. Matching concept is assessed into two classes like Syntax and linguistics Structures. Syntax matching concept is mainly focuses on syntax supported to the acceptable compiler. Linguistics is the main accustomed resolve the given word victimization logical analysis. The main objective of this proposed work is to determine the probability of semantic word used in the e-content which is retrieved from the given document. These techniques used to stem and trim the word from the given document and classify based on the knowledge such as Factual, Procedural and Conceptual. These Classified words are reconstructed into tree structures, used to calculate the probability of outcome and evidence. These effective and effusive techniques mainly to reduce the time, memory utilization and efficiency based on the proposed SAS (Semi-Automated Semantic) algorithm.

Keywords: Semantic, Syntax, ontologies, Structures, stem.

I. INTRODUCTION

The integration of ontologies is a major challenge and research issue in semantic web. Such as finding similarities and difference among ontologies in automatic and semi-automatic way, defining mapping between ontologies, composing mappings across different ontologies has to be faced during managing these diverse ontologies. Ontology management is possible through interoperability of semantic data sources [1]. The semantic technologies such as XML and ontology can play important role for the development of semantic based information retrieval. It supports more expressive queries and produce accurate results for this we have collected the documents from the different domains and design the tree structures of the documents in the form of xml and ontology and data mining technique such as clustering and then retrieve the information from this structure and based on user interest that provide the concept based [2].

The development of linguistics internet technologies has been closely associated with the planet wide internet. This is providing the artificer of the WWW – Sir Tim Berners-Lee has originally coined the term “Semantic Web” and has impressed a lot of analysis during this space [4]. And almost like the normal internet, the inspiration of semantic internet technologies square measure information formats which will be accustomed cipher information for the process (relevant aspects of it) in laptop systems. However, viewing the WWW because the sole origin and inspiration for the technologies that square measure delineate during this study wouldn't do justice to their true history. Additional significantly, it’s not conjointly hiding a number of the most motivations that have semiconductor diode to the technologies in their gift type.

II. RELATED WORK

The general approach of one is building abstract models that capture the complexities of the planet in terms of easier concepts. Modelling during this sense pervades human history – a comprehensive historical account is on the far side the scope of this study – however underlying strategies and motivations square measure extremely relevant for the linguistics technologies that square measure out there for the USA these days. A second, the moremodern approach is that the plan of computing with information [4]. The vision of representing information in an exceedingly means that permits machines to mechanically return to cheap conclusions, perhaps even to “think,” has been a propulsion for many years of analysis and development, long before the WWW was unreal.

The linguistics internet has been planned as an associate extension of the planet wide internet that permits computers to show intelligence search, combine, and method web page supported which means that this content should humans [5]. Within the absence of human-level AI, this could solely be accomplished if the supposed which means (i.e. the semantics) of internet resources is expressly per a format that's process able by computers. For this it's not enough to store information in an exceedingly machine-process able syntax – each mark-up language page on the net is machine-process able in an exceedingly sense – however, it's conjointly needed that this information is blessed with a proper linguistics that clearly specifies that conclusions ought to be drawn from the collected info. Clearly, this might be associated not possible endeavour once aiming in the least human information found on the net, providing it's usually exhausting enough for humans to even agree on the contents of a definite document, to not mention formalizing it in an exceedingly means that's purposeful to computers.

III. METHODOLOGY

3.1 SEMI-AUTOMATIC SEMANTIC CONCEPT

This Matching idea is assessed into two classes like Syntax and linguistics Structures. Syntax matching idea in the main focuses on syntax supported the acceptable by compiler. Linguistics is the main accustomed resolve the given word victimization logical analysis. As an example, a user within the geographical area could use some technical words compare to a country [10]. This Classification of words referred to the linguistics. A linguistics Structure is categorized into two parts as shown within the figure 2.1 component level and structure level. component Level that determines the content on a component basis like for bookstore: book id, book name, book author, book ISSN, book price and book edition of all attributes square measure in the main targeted on parts. Next is structure levels, that square measure classified into four technical terms such as a) Ontology b) Semantic c) Tree and d) Iteration.

a) **Ontology** encompasses associate illustration, formal naming, and definition of the categories, properties, and also the relation between the concepts, information, and entities that substantiate one, many or all domains [5][10]. Every field creates ontologies to limit the standard and organize information into info and data. As new ontologies square measure created, their use hopefully improves downside finding at intervals that domain [9].

b) **Semantic** is that the sphere committed the rigorous mathematical study of the suggest that of programming languages. It'll thus by evaluate which suggests of syntactically valid strings printed by a specific language, showing the computation involved. In such a case that the analysis would be of syntactically invalid strings, the result would be non-computation.

c) **Tree** is commonly printed recursively (locally) as a gaggle of nodes (starting at a root node), where each node is also a company consisting of a worth, in conjunction with an inventory of references to nodes (the "children"), with the constraints that no reference is duplicated, and none points to the inspiration.

d) **Iteration** is that the act of repetition a way, to return up with a (possibly unbounded) sequence of outcomes, with the aim of approaching a desired goal, target or result. each repetition of the tactic is to boot brought up as degree "iteration", and so the results of one iteration square measure used as a result of the beginning line for consecutive iteration.

A Combination of linguistics and Tree Structure that ends up in Semi-automatic linguistics Matched construct.

The Semantic Web offers the possibility of providing the meanings or semantics of web documents in a machine-readable manner. However, the vast majority of 1.5 billion web documents are still in a human-readable format, and it is expected that this form of representation will still be the choice among content creators and developers due to its simplicity. Due to this phenomenon and the desire to make the Semantic Web vision a reality, two approaches have been proposed either furnish information sources with annotations that provide their semantics in a machine-accessible manner or write programs that extract such semantics of web sources[6].

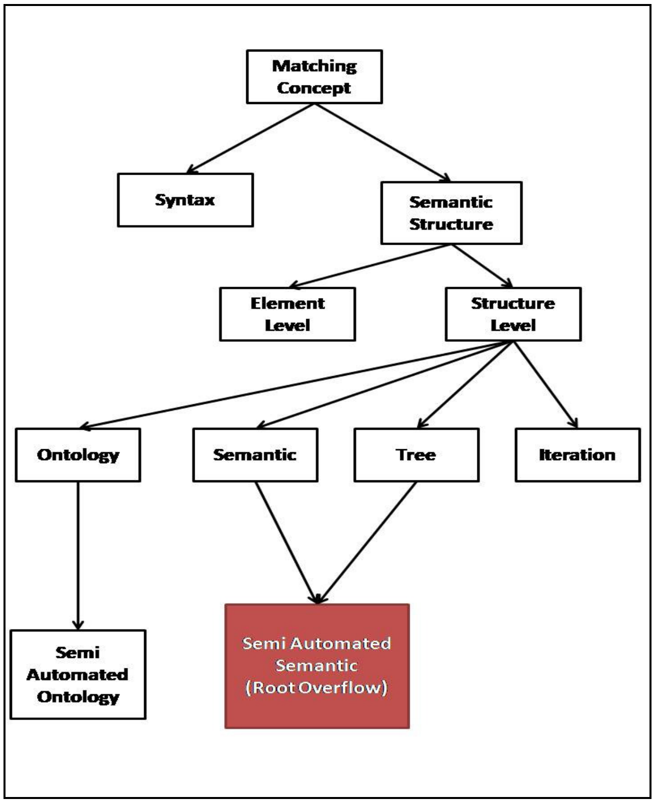

Figure 3.1 Structure of Matching Concept

3.2 Algorithm for SAS (Semi-Automatic Semantic)

In SAS, the Probability of having both the Outcome O and Evidence E is: (Probability of O occurring) multiplied by the (Prob of E given that O happened).

The evidence, P(Outcome or Evidence) = P(Evidence given that the Outcome) times Prob(Outcome), scaled by the P(Evidence).

In Naive Bayes, to predict an outcome of multiple evidence that case, the math gets very complicated. To get around that complication, one approach is to 'uncouple' multiple pieces of evidence and to treat each of pieces of evidence as an independent[11]. This approach is called SAS.

P(Outcome/evidence) = (P(Likelihood of Evidence) x Prior prob of outcome)/ P(Evidence)

The intuition behind multiplying by the prior is so that gives high probability to more common outcomes and low probabilities to unlikely outcomes. These area unit known as base rates and that they are the way to scale our foretold chances. The formula above for each possible outcome is trying to classify, each outcome is called a document and it has a document label. The job is to look at the evidence, to consider how likely it is to be this document is assigned a label to each entity. Processes of SAS are as follows:

In this algorithm, it mention D as files, N as no of words, C as Evidence, T as outcome, ST as list of stopwords, L as list of files, V as vocabulary.

```
Files(C, D)
            ST <- LoadStopwordList
            L <- documentList
            IF ST is empty THEN
                FOR each W in ST LOOP
                        L <- removestopword(W)
                Next LOOP
            End IF
            IF L is empty THEN
                FOR each K in iterate (L) LOOP
                        S <- K in documentList(L)
                        J <- processthestemming(S)
                        Add to J to stemmingList(J)
                Next LOOP
                V <- ExtractVocabulary(J)
                    N <- Countword(J)
                    FOR each c with J LOOP
                    do N^c <- countwordinDoc(J, C)
            prior[c] <- N^c/N
                text <- concatenateTextofallWordAsDoc
FOR each t with V LOOP
do T <- countwordOfTerm(Text, t)
            FOR each t with V LOOP
            do condprob[t][c] <- (T+1)/sum(T + 1)
                                Next t in LOOP
                        Next t in LOOP
                    Next c in LOOP
                End IF
    return V, prior, condprob
```

IV. RESULTS AND DISCUSSION

Consider D as given document, N be the number of word in the given document. Initially the given document D is used for Stemming process. The Process which is used to eliminate the unwanted word based on Blooms taxonomy. Next process is of trimming which is used to remove the prefix and suffix of the adjective word from the given document D. Consider the Sample data for Knowledge extraction such as Factual knowledge, Procedural Knowledge and Conceptual knowledge. The probabilities of outcomes are calculated and compare with the existing algorithms.

As implement an algorithm, A Sample document consist of 256 words, the following table 4.1 are generated based on the SAS Algorithm.

Table 4.1 Probability of Outcome using SAS

Knowledge Representation	Total No. of Knowledge Words	Total .No.of Words in the Document	Probability (Likelihood of Evidence)	Prior Probability of Outcome	Probability of Evidence	Probability of Outcome
FACTUAL	21	81	0.05	0.32	0.08	0.18
PROCEDURAL	48	98	0.02	0.38	0.19	0.04
CONCEPTUAL	35	69	0.03	0.27	0.14	0.06

After stemming, trimming, calculate the number of words, probability (likelihood of evidence), Prior of Probability of outcome, probability of evidence to determine probability of outcome. Based on the outcome, maximum (P (outcome)) is knowledge of the appropriate document. From the table 4.1. The given document mainly focused on Factual of Knowledge.

Based on the above table 4.1 the following graph Figure 4.2 generated as follows to determine the knowledge of particular document

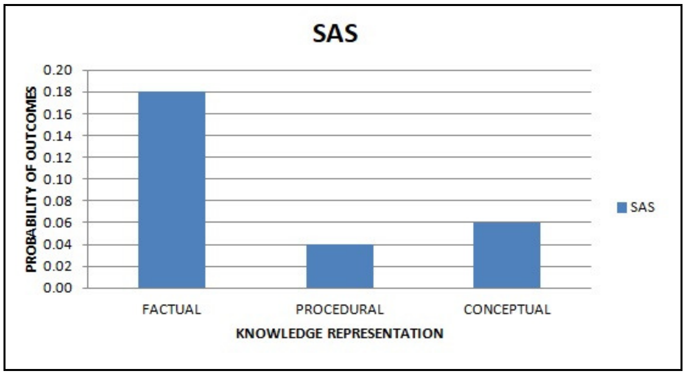

Figure 4.2 Comparison of Knowledge on SAS

V. CONCLUSION AND FUTURE SCOPE

SAS uses a file containing the grammatical rules for languages encoded in a standardized format and a dictionary file containing the languages valid stems. The analysis of a word loops over the grammatical rules applying those applicable and then checking if the valid stem is found. The advantages of SAS are that very intricate grammatical rules can be applied such as the removal of multiple suffixes and prefixes. Suffix stripping algorithm may differ in the result for a variety of reasons. One such reason is the algorithm constrains the output word must be a real word in the given language. Future work may lead to big data, since the number of users may increase in terms of percentage. This may be extended to n-tier architecture. This mobile application may be developed and designed in the swift programming, where the application may run on any operating system such as iOS, Blackberry and Symbian operating system.

REFERENCES

[1] Ranjna Jain, Neelam Duhan, A.K.Sharma,"Comparative Study on Ontology Management Approaches in Semantic Web",IJCSE Vol 6, Issue 1, PP 132-140. -Jan 2018.

[2] Muqeem Ahmed,"Semantic Based Intelligent Information Retrieval through Data mining and Ontology", IJCSE VOL- 5, ISSUE - 10,PP 210-217,Oct-2017.

[3] S. Banerjee,"A Semantic Web Based Ontology in the Financial Domain", International Journal of Computer and Information Engineering, Vol:7, No:6,PP 807-810, 2013.

[4] H. Srimathi, "Semantic Web based Personalized eLearning", International Journal of Applied Information Systems (IJAIS), Volume 2– No.1, PP 11-16, May 2012.

[5] Vasani Krunal A, "CONTENT EVOCATION USING WEB SCRAPING AND SEMANTIC ILLUSTRATION",IOSR Journal of Computer Engineering (IOSR-JCE),Volume 16, Issue 3, Ver. IX , PP 54-60,May-Jun. 2014.

[6] Hasida, K. "Semantic Authoring and Semantic Computing". Sakurai, A. et al. (Eds.): JSAI 2003/2004, LNAI 3609, PP 137–149.

[7] Diana Man, "Ontologies in Computer Science, Didactica Mathematica", Vol. 31, No 1, pp. 43–46,2013.

[8] YassineGargouri,"Ontology Maintenance using Textual Analysis, Systematic-cybernetics and informatics",Vol-1,Number 5,pp-63-68,2016.

[9] Arun K.Pujari,, ―Data mining techniques‖, Universities Press (India) Pvt. Ltd. 2001.International Journal of Management, Technology And Engineering, Volume 8, Issue IX, SEPTEMBER-2018.

[10] Kristina M Doing-Harris, "Automated Concept, and Relationship Extraction from the Semi- Automated Ontology Management (SEAM) System", International Journal Biomedical Semantics, April 2015.

[11] D. Elangovan, "Semi-Automated Semantic Matched Concept Extraction Model for E-Content Development", International Journal of Applied Engineering Research, November 5, 2016.

Authors Profile

First Author: D.ELANGOVAN. MCA, MPHIL, SET .,Research Scholar, Manonmanium Sundaranar University, Thirunelveli. Working as a Assistant Professor in D.B.Jain College, Chennai – 97. Worked as a senior software Engineer in CGI Bangalore, Member in ICT ACADEMY OF TAMIL NADU, ORACLE ACADEMY, Selection Committee Member in Government college of fine Arts,Chennai-india.

Second Author: Dr.K.NIRMALA, M.C.A.,Ph.D., *Research supervisor* Manonmanium Sundaranar University, *& Associate professor, Department of computer science, Quaid-e-millath govt.college for women Chennai-02, selection committee member TRB.Research supervisor in various university.*