ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Machine Learning Approach for Self Adaptive Semantic Focused Crawler Based Data Mining Services

Gajanan V. Jaybhaye, Prof A.V.Deorankar

M.Tech Student, Dept. of CSE, Government College of Engineering, Amravati, India

Associate Professor, Dept. of I.T., Government College of Engineering, Amravati, India

ABSTRACT: Crawling is one of the important systems for building knowledge stockpiles. Focused crawling is aimed at specifically finding out pages that are pertinent to a predefined set of subjects. The cause for semantic focused crawler is naturally finding, formatting and ordering the administration data with the semantic web advances. Heterogeneity, universality and equivocalness are the three major problems with administration clients when searching for mining administration data onto the internet. In this paper, we present the structure of a new self-adaptive semantic focused crawler with machine learning approach, the motive of definitely and proficiently finding, arranging and indexing mining administration data onto the internet, with high performance rate by taking into account the three noteworthy issues. This structure assembles the technologies of semantic focused crawling and machine learning, in order to nurture the performance of this crawler, heedless of the variety in the web environment. Also it uses the concepts of word net and semantic similarity.

KEYWORDS: Machine Learning; Semantic Similarity; Service Advertisements; Prediction; Semantic Focused Crawler.

I. INTRODUCTION

It is decently perceived that data innovation has a significant impact on the way business is led, and the Internet has turned into the biggest commercial centre on the planet. Inventive business experts have understood the business applications of the Internet both for their clients and key accomplices, transforming the Internet into a colossal shopping centre with a colossal inventory. Purchasers have the capacity scan a tremendous scope of items and administration notices over the Internet, and purchase these products specifically through online exchange frameworks. Administration notices structure an impressive piece of the publicizing which happens over the Internet and have the accompanying peculiarities [2] [3] [4].

When we search any data on the internet there are number issues occurs at the time of searching, these issues includes- Heterogeneity, Universality and Equivocalness lets see these issues one by one.

A. Heterogeneity

There are number of ways to upload services over internet with multiple domain. This services can be classified based on ownership, demand, supply and the impact of service [5] [6]. But there is not an agreed upon approach to classify this services.

B. Universality

Administration promotions can be enrolled by administration suppliers through different administration registries, counting-

1.Worldwide business internet searchers, for example, Business.com.

2. Neighbourhood professional resources, for example, Google Local Business Centre and nearby Yellowpages5.

3. Space particular business web search engines, for example, medicinal services, industry and tourism business web search engines.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

4. Web search tool publicizing, for example, Google6 and Yahoo!7 Advertising Home[7]. These administration registries are topographically circulated over the Internet.

C. Equivocalness

Amount of information present over the internet is described in natural language therefore it may be unclear. Moreover, online service information does not have a coherent format and standard, and differs from Web page to Web page. Mining is one of the olden industries in human history, having appeared with the beginning of human civilization. Mining services refer to a series of services which aid mining, quarrying, and oil and gas extraction activities. Since the arrival of the information age, mining service companies have bethink the power of online advertising and they have endeavour to promote themselves by actively joining the service advertising community [8].

In order to address the above problems the framework of a novel self-adaptive semantic focused (SASF) crawler, by mixing the technologies of semantic focused crawling and machine learning is design, whereby semantic focused crawling technology is used to resolve the issues of heterogeneity, universality and equivocalness of mining ploy information, and machine learning technology is used to nurture the high performance of crawling in the uncontrolled network environment. This crawler is designed with the motive of helping search engines to precisely and capable of search mining service information by semantically finding, arranging, and indexing information.

Apart from this, here we are using here the concept of semantic similarity and word net, semantic similarity which is the idea of distance between them is crutch on the likeness of their meaning or semantic content as defy to similarity which can be estimated regarding their syntactical representation and word net is a lexical database for the English languages.

Also, here we are using machine learning , it is a subfield of computer science that develops from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning traverse the study and construction of algorithms that can learn from and make prediction on data.

II. RELATED WORK

A semantic focused crawler is a software agent that is used to cross the web and retrieve as well as download related web information on specific topics by means semantic technologies[9].

H. Dong et al.[1] proposed a self adaptive semantic focused crawler for mining services information discovery. It is based on ontology learning approach. It uses the ontology as repository and generate the metadata.It has drawback regarding the performance of the self adaptive model did not completely meet expectations regarding the parameters of precision and recall. W. Wong et al.[10] proposed a crawler in which attention is towards the enhancing semantic focused crawling technologies by combining them with ontology learning technologies. It contains drawback ralating to the differentiation and dynamism. Zheng et al.[11] proposed a supervised ontology learning based focused crawler that aims to maintain the harvest rate of the crawler in the crawling process. The prime idea of this crawler is to construct an artificial neural network model to determine the relatedness between a web composition and an ontology.It does not have the function of classification. It cannot be used to develop ontologies by enriching the vocabulary of ontologies. The supervised learning may not work within an uncontrolled network environment with unpredicted new terms.

C. su et al.[12] proposed an unsupervised ontology learning based focused crawler in order to enumerate the relevance scores between topics and web documents. Given a specific domain ontology and a topic represented by a concepts in this ontology, the relevance score between a web documents and the topic is weighted sum of the befall frequencies of all the concepts at the ontology in the web documents. Also this crawler makes use of reinforcement learning, which is probabilistic framework for learning optimal decision making from rewards or punishments [13], in order to train the weight of each concept. It has drawback like, it cannot be used to enrich the vocabulary of ontology.

III. PROPOSED WORK

In proposed work section, we will prefaces the system workflow of the SASF crawler step by step as shown in fig 1. The goals of this crawler include- to generate mining service metadata from web pages and to exactly associate between the semantically pertinent mining service concepts and mining service metadata with relatively low computing cost.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

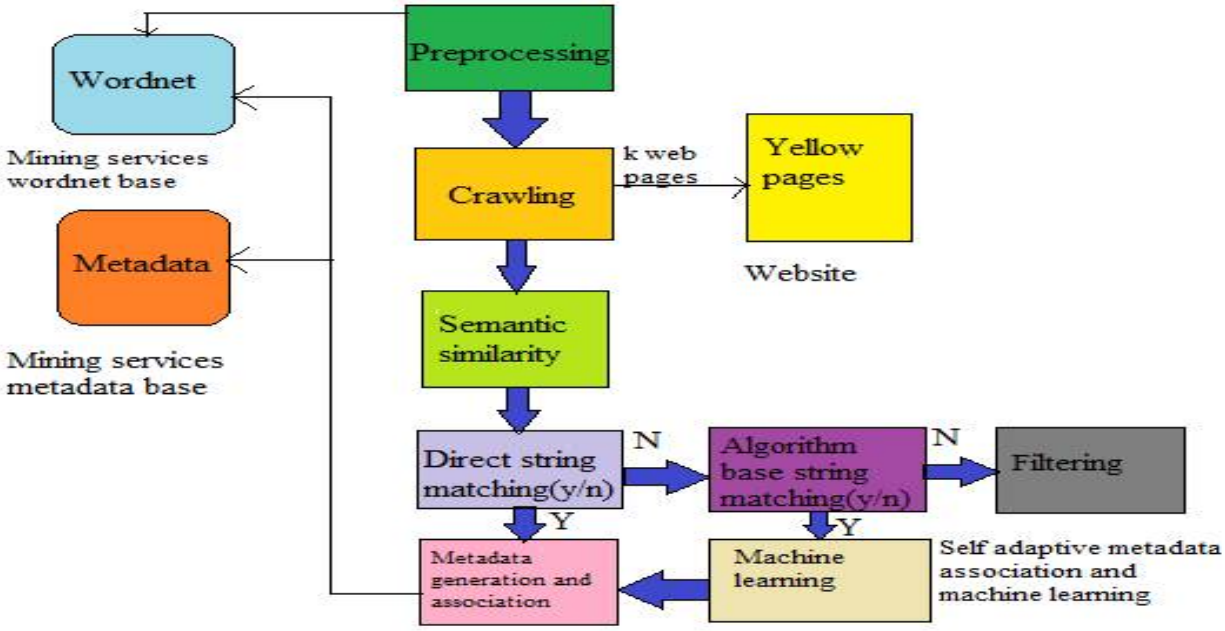

Fig 1: System architecture of the proposed self adaptive semantic focused crawler

In fig 1 system architecture of the proposed self adaptive semantic focused crawler is shown it is based on the machine learning approach. Here, design a SASF crawler with machine learning with direct string matching, algorithm base string matching, metadata generation and association and filtering.

The first step is pre-processing , which is to process the concept in the word net before matching the metadata and the concepts.

The second step is crawling in which it will download k web pages from the internet at one time and to concentrate the needed in-arrangement from the downloaded web pages as in indicated by the mining administration metadata blueprint and the mining administration supplier metadata mapping, keeping in mind the end goal to set up the property estimations to produce another gathering of metadata.

The next step is semantic similarity which is the idea of distance between them is crutch on the likeness of their meaning or semantic content as defy to similarity which can be estimated regarding their syntactical representation . It is used find out the semantic similarity between two sentences.

The rest of the system workflow can be integrated as a self adaptive metadata association and machine learning process. The details as follows- first of all, the direct string matching process inspects it will check the metadata that are already present in word net with the new concepts. If the solution is 'yes', then the concept and the metadata are considered as semantically pertinent by means of metadata generation and association process the metadata to be able to produced and stored in mining service metadata base as well as being associated with the concept. Suppose, the answer is 'no' an algorithm basis string matching will be invoked to search the semantic relatedness between the metadata and the idea by means of a concept metadata semantic similarity algorithm. If the concept and metadata are semantically pertinent, then the metadata can be considered as a novel value for the concept. It can be done using the machine learning approach and the metadata is thus allowed to go utter the metadata generation and association process, or else the metadata is considered as semantically non pertinent to the concept. The above process is repeated till all concepts in the mining service word net have been paragon with the metadata and suppose, the none of the concepts is semantically pertinent to the metadata this metadata is considered as semantically non pertinent to the mining service domain and will be filtered out.

The concepts can be utilized in the algorithm based string matching process because the semantic relatedness between the concept and metadata is calculated by comparing their algorithm based similarity values with the threshold

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

value. If the majority similarity value between metadata and the concept is higher than the threshold value, the metadata and the concept are considered as semantically pertinent; otherwise not [1]. By the use of machine learning we can increase the performance rate of the crawler in uncontrolled network environment.

A. *Mining Service Word Net Base and Mining Service Metadata Base*

Word Net is a lexical database for the English language. It groups English words into sets of synonyms, provides small definitions and usage examples, and records a number of relations among these synonym sets of their members. Word net can thus be seen as a synthesis of dictionary and thesaurus. Its primary use is in automatic text analysis and artificial intelligence applications.

Mining Service Metadata Base is used to store the automatically produced and indexed mining service metadata.

B. *Semantic Similarity*

In Semantic Similarity approach, the meaning of a target text is inferred by checking how similar it is to another text called the benchmark text, whose meaning is known. If the two texts are similar enough, in conformity with to some measure of semantic similarity, the meaning of the target text is deemed similar to the meaning of the benchmark text. It is used to find out the semantic similarity between the two sentences [14].

C. *Machine Learning*

Machine Learning is a sub domain of computer science that evolved from the practice of pattern recognition and computational learning theory in artificial intelligence. It performed prediction on data by using some algorithm. Also it focuses more on exploratory data analysis. Machine learning tasks include- unsupervised learning, supervised learning and reinforcement learning.

D. *Linguistic-Based Chain Matching Algorithm*

The key thought of the Linguistic-Based Chain Matching calculation is to gauge the content closeness between an concept description property of word net and an service description property of metadata, by means ofwordnet9 and a semantic comparability model. As the concept description and the service description can be viewed as two gatherings of terms after the pre-processing and term handling stage, most importantly, we have to look at the semantic comparability between any two terms from these two gatherings. Since terms
(On the other hand ideas) in Word net are sorted out in a various levelled structure, in which concepts have the connections of hyponym, it is conceivable to survey the closeness between two ideas by looking at their relative position in Word net. Resniks model can be communicated as

$$resem_{Resnik}(P1, P2) = \max_{p \in t(p1,p2)} [-\log (D(P))]$$

Where P1 and P2 are two concepts in Word Net, and T(P1, P2)is the set of concepts that subsume both P1 and P2 and D(P) is the probability of encountering a sub-concept of P Hence,

$$D(P) = \frac{d(P)}{\ominus}$$

Where d(P) is the number of concepts subsumed by P and $\ominus$ is the total number of concepts in Word Net [14]. Result of Resniks model is within the interval [0, ∞].

eq. (3)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

IV. IMPLEMENTATION

In this section, we have design one module, in which we have design one web crawler which used to download the relevant web information. Web crawler contains the URL and GO button to retrieve the information. Fig 2 shows the web crawler, in which we have to enter the URL. After entering the URL, click the GO button, it has shown in fig 2.

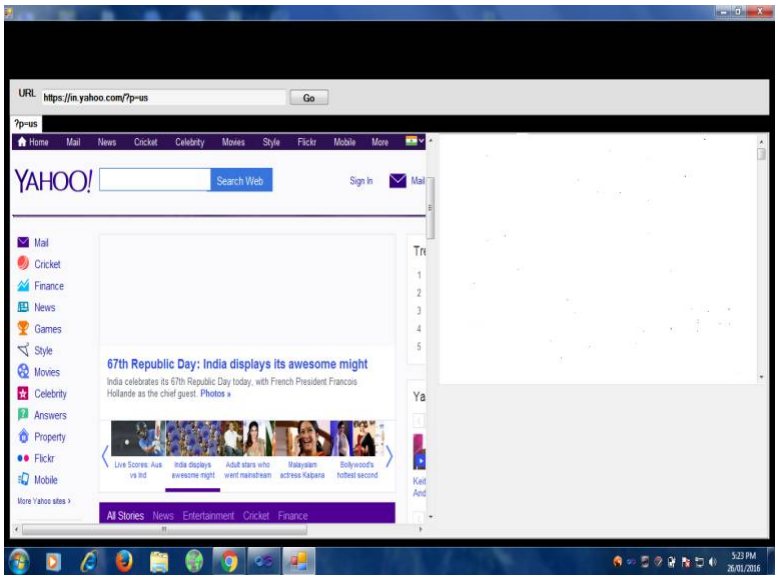

Fig 2 Screenshot of web crawler page

Fig 3 shows the screenshot of web crawler page, in which we have to enter URL and after clicking on GO button, it will shows the downloaded page in left side panel, there is two panel, one for to shows the downloaded page and other for to shows the content descriptions of that page. Content descriptions will be beneficial for the metadata generation and association process. Up to this we have done our project and it is in progress.

We have just design a basic structure of the crawler for which it will only crawl the web pages, after downloading the pages it will mine some kind of services e.g. fruits services, mobile services, hotel services or any kind of services it will mine for which our crawler gives the fast access to this services for user benefit.

In this project we using the Word net dictionary, in which metadata are stored, it is predefined dictionary and newly created metadata are also stored in it, when we entered some word it will rank the documents on the basis of word present in it.

First of all the word which are entered by user on which direct string matching is done if the word already present in word net it provide direct access to that word otherwise that word is check by algorithm that we have discuss previously.

Algorithm will check all the relevant data that are present in word net, if it is match then new metadata are generated and which will stored in mining services metadata base and further used by the mining services word net base. If the relevant data will not occurs then the term will be filter out.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

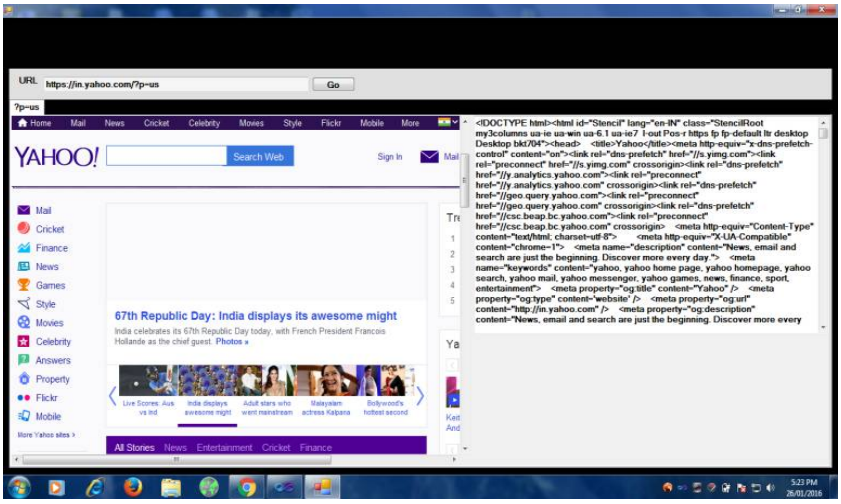

Fig 3 Screenshot of web crawler page

V. CONCLUSION

Here, we have proposed an innovative machine-learning basis focused crawler, by taking into account these three problems i.e. Heterogeneity, Universality and Equivocalness that available onto the internet. This approach involved a novel concept metadata matching algorithm. Here, we have presented Linguistic based chain matching algorithm to find out relevant concepts. Also we have presented here our progress work, in which we have design one web crawler for efficient retrieval of information. The algorithm we have presented here it will used to enhance the performance of web crawler.

REFERENCES

1. Hai Dong, member, IEEE, and Farookh Khadeer Hussain, "Self Adaptive Semantic Focused Crawler for Mining Services Information Discovery" IEEE Transactions on Industrial,Informatics,vol.10,No.2, pp.1616-1626, May 2014.
2.B. Fabian, T. Ermakova, and C. Muller, "SHARDIS A privacy-enhanced discovery service for RFID-based product information", IEEE Trans. Ind. Informat., to be published.
3. M. Ruta, F. Scioscia, E. D. Sciascio, and G. Loseto, "Semantic-based enhancement of ISO/IEC 145433 EIB/KNX standard for building automation", IEEE Trans. Ind. Informat., vol. 7, no. 4, pp. 731739, Nov. 2011
.4. I. M. Delamer and J. L. M. Lastra, "Service-oriented architecture for distributed publish/ subscribe middleware in electronics production", IEEE Trans. Ind. Informat., vol.2, no. 4, pp. 281294, Nov. 2006.
5. H. Wang, M. K. O. Lee, and C. Wang, "Consumer privacy concerns about Internet marketing", Commun. ACM, vol. 41, pp. 6370, 1998.
6. R. C. Judd, "The case for redefining services", J. Marketing, vol. 28,pp. 5859, 1964.
7. H. Dong, F. K. Hussain, and E. Chang, "A service search engine for the industrial digital ecosystems", IEEE Trans. Ind. Electron., vol. 58,no. 6, pp. 21832196, Jun.2011. |
8. Mining Services in the US: Market Research Report IBIS World 2011.
9.H. Dong and F. K. Hussain, "Focused crawling for automatic service discovery, annotation, and classification in industrial digital ecosystems," *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2106–2116, Jun. 2011.
10. W. Wong, W. Liu, and M. Bennamoun, "Ontology learning from text: A look back and into the future," *ACM Comput. Surveys*, vol. 44, pp. 20:1–36, 2012.
11. H.-T. Zheng, B.-Y. Kang, and H.-G. Kim, "An ontology-based approach to learnable focused crawling," *Inf. Sciences*, vol. 178, pp. 4512–4522, 2008.
12. C. Su, Y. Gao, J. Yang, and B. Luo, "An efficient adaptive focused crawler based on ontology learning," in *Proc. 5th Int. Conf. Hybrid Intell. Syst. (HIS '05)*, Rio de Janeiro, Brazil, pp. 73–78,2005.
13.J. Rennie and A. McCallum, "Using reinforcement learning to spider the Web efficiently," in *Proc. 16th Int. Conf. Mach. Learning (ICML '99)*, Bled, Slovenia, pp. 335–343, 1999.
14. P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *J. Artif. Intell. Res.*, vol. 11, pp. 95–130, 1999.