Abstract: Location information are important in location sensitive tasks. Web pages contains large number of location information. Web pages are often noisy, massive. This paper propose a method to find the location information from web pages and also here NER (name, entity, recognition) method is implemented. By using NER we can predict whether the search query contains persons, organization and location information. Here considers both location terms and non-location terms in a web page. The number of occurrences of word is counted to find the weight. It is always difficult to consider all terms or datas in web page so that here only metadata content is taken to avoid complications. tfidf algorithm is used to avoid noisy datas or massive ones.

Date of Submission: 30-04-2019
Date of acceptance: 15-05-2019

I. Introduction

Now people are increasingly using internet and also different applications of internet are available. Now there are lot of applications are available to find the location of a person. But now the techniques to find the location information of web pages are limited and it has its own drawbacks. For eg; If a person wants to know more about a school which is nearer to his current location. If he searches just ‘schools’ only. He is expecting schools nearer to his current location. Now so many applications are available to find the users location i.e GPS. The location information about web pages are limited. Here this paper proposed a method called NER to identify the Location, Person and organization within the search query itself. Here introduced an algorithm to find the location names in a web page. Content information of web pages are often noisy and massive. It may contains URLs, titles, abstracts, main body contents etc. Number of web pages are increasing day by day. Some web pages are using web pages for different purposes. Someone will use implicit search queries, whereas others use explicit search queries. Some methods had used natural language processing and machine learning techniques[2][3][4][5][6]. Although the potential location information in non location terms are ignored. Here this paper considers both location terms and non location terms and search queries. We will find how many times that the location is displayed in that particular meta data content. In twitter data we have location-tagged tweets but there will be no information about location of web pages. The stopwords are removed by using the steps stemming rooting etc.

II. Framework And Methods Overview

This prospective comparative study was carried by using an algorithm and a method called NER. The framework contains training section and testing section. The search queries or the location prediction is directly connected with the google. It will display the web pages related to the search queries and may extract both location terms and non location terms from the web page content. The meta data content will be taken out from large data so that it can avoid the noise. The location names will extract from the metadata content and will find how many times the location names are present in meta data content. The next section is testing section where will test the contents and will predict the locations. tfidf algorithm is implemented to avoid the noisy and massive data. NER recognizes the location, person and locations in a particular query. tfidf can predict how important a word in a document. this will make you more knowledge about the important word and their importance in document. It has more advantages than any other algorithms. Considering all datas in a website is a biggest challenge we are facing. This algorithm and methods has a great value by considering all terms including non location terms. Instead of considering data which is massive we will only consider the main terms. No extra effort is needed to extract and weight the terms. This methods can avoid the disadvantages of previous algorithm where twitter data or location- tagged tweets and users profile informations are available to find the locations. Large scale location prediction is found by using tfidf algorithm by taking a large website. A novalable and scalable framework is implemented to find the locations. The stopwords are removed by using the steps stemming rooting etc so as to avoid complications. The links are extracted from the websites by using the link extraction methods. Term vector weight will indicate the importance of that term in that data which is present in the website. Number of occurences of main words is denoted as a count to predict the weight of that word in a
Location extraction is also finds the number of occurrences of location in the metadata content of the website. A simple framework is demonstrated here to denote the location extraction from a large data.

**Training Section Overview**

```
Search Queries

Web Pages

Location Terms
Non location Terms

Metadata Content

Find The Location Terms

Display Location Names (with the number of times it is present in metadata content)
```

**Testing Section Overview**

```
Test Content

Predicted Locations
```
NER

Major Terms Used In This Paper
1. Term: A word, it can be single or multiple
2. Location Terms: Terms that are related to location names eg: streetnames, cities.
3. Non location terms: Terms that are not related to location terms.
4. Implicit search query: query without explicit information about location.
5. Explicit search query: query with explicit information about location.
6. NER: Name Entity Recognition.

Algorithm: To find the location vectors
For url u do
Initialize weight of term=0
For each query q which is connected to u do following step
For each term v in q
    Update weight
Normalize weight

Location-URL click graph
If q is location sensitive query
If query is explicit query
Extract location from query
If query is implicit query
Extract location of users as query information.

Term location vector
For each term location
Initialize term weight=0
For each location terms
Update location terms weight

NER
It is name entity recognition of search query. It is connected to the Wikipedia.
For each term t in search query
Check t is a person, organization or location
If the term t is about personal information
It is a Person
If the term t is related to organization
It is Organization
If the term is about street names, cities etc. It is Location.

**Algorithm: TFIDF**

Let $t$ be a term in document $d$, weight $w_{t,d}$ of term $t$ in document $d$ is $w_{t,d}=\frac{TF_{t,d}}{DF_{t}} \log \frac{N}{DF_{t}}$

$TF_{t,d}$ is the number of occurrences of $t$ in document $d$, $DF_{t}$ is the number of documents containing the term $t$, and $N$ is the total number of documents.

The contents, main bodies etc are all adopted in location prediction. It is always a difficult to consider all data so here only metadata content was taken, even if we search a query a query URL click graph is formed. Each term can be denoted as a vector. Weight of each vector is directly proportional to the confidence of term related to location.

**III. Conclusion**

Here this paper proposed an efficient algorithm to find location of web pages. Here also implemented NER (name entity recognition) to find the name, organization and location details within the search query itself. It considers all the location terms and non-location terms within a web page. It is entirely different from the traditional methods which focus on search query only. TFIDF algorithm is used to filter out the unwanted noisy and massive data.

**References**

[1]. Yueninghu, Changsungkang, Jiliangtang, Dawei yin and Yi chang Large scale location prediction for web pages program September 2017
[5]. Qin, R xiao, L fang, x xie and l zhao “an efficient location extraction algorithm by leveraging web contextual information.” In proc ACM intsym advances geographic infosyst 2010, pp 53-60
[7]. H Deng, L king and M R lya “entropy-biased models for query representation on the click graph” in proc ACM SIGIR conf res develop infreieval 2009, pp 339-346
[8]. HL wang et al “detecting dominant locations from search queries” in proc ACM SIGIR conf res develop infreieval, 2005, pp424-431