Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 04, April -2018

SENSE BASED INDEXING OF HIDDEN WEB USING ONTOLOGY

Sushmita singh

YMCAUST, Faridabad

Abstract:- *Today when internet is considered to be a storehouse of information where one can get everything and every kind of information one needs. Thus with growing world and the growing needs of people the volume of information on internet is increasing tremendously. To handle this large amount of information, Web searchers use search engines. But what we browse through usual search engines is only the surface web. But there also exists a hidden web which contains a large collection of data that we cannot access just by using normal search engines which are solely based on hyperlinks. This hidden data is accessed by passing through different types of barriers generally it is a form based gateway. After filling in the form one gets the entrance in the hidden web. However indexing helps in arranging the search or the query results without bypassing this form. Different techniques have been developed for this purpose. The main objective of this paper is to find out a method that indexes the hidden web pages according to its relevance for a particular context of a search query, a method that uses the prevailing techniques and gives the required output using ontology. One of the advantages of storing an index is that it optimizes the performance and also it finds out relevant documents from the search engine storage area for a user given search query with specific context.*

Keywords: *Hidden web, indexing, ontology, homonyms, attribute extraction.*

1. Introduction

To access the hidden part different hidden web crawlers have been developed and different methodologies have been proposed. Generally the barrier between the user and the hidden web is a form that needs to be filled by the user sometimes for authentication and sometimes to limit the scope of the search. After filling this form the resultant web pages from the previous step are stored and then to improvise the search results for a user query we need indexing techniques.

While indexing a common problem faced is the problem with homonyms (single word with multiple contexts and meanings). This issue needs to be resolved so as to clarify with the context of the webpage that we are indexing that is to identify correctly to which context the webpage belongs. Different techniques have been evolved for solving this problem with multiple meanings. Word sense disambiguation is an open problem of NLP (natural language processing) which eliminates the ambiguity between multiple meanings of a single word. This helps in unfolding the context of the webpage. The main task is to reduce the complexity while side by side increasing the efficiency and mainly providing clarity with the context of the webpage.
This paper is organized as follows Section 2 briefly describes the basic concepts that are related to this algorithm, Section 3 presents the overview and architectural design of the proposed algorithm, Section 4 explains the proposed algorithm, Section 5

2. Basic concepts

2.1 Hidden web:

The World Wide Web is divided into two parts publically indexed web also called surface web and the hidden web or the deep web. More than 90% of the total web data is hidden. The Deep Web refers to Web data sources that provide a considerable amount of information with backend databases that are not indexed by general search engines. The Deep Web sources require manual query interfaces (typically appearing as Web form pages) and dynamic programs to access their contents, thus preventing Web crawlers from automatically extracting their contents and indexing them, and therefore not being included in search engine results. Routine crawlers are not that successful with this part of the web thus we require special hidden web crawlers for the deep web.

2.2 Attribute extraction:

An attribute is something that defines a quality or characteristic of a particular thing. Attribute in our context may be defined as a basic element of the web page that describes the domain of the web page. Attributes of a webpage are those words on which one can rely for the context of the webpage. Automatic attribute extraction focuses on discovering the importance of web page according to the context based on the keywords. Extracting attributes helps in unveiling the context of the webpage and hence can prove to be an important step in indexing of the deep web.

2.3 Ontology (WordNet):

Ontology is a formal representation which includes vocabulary for referring to the terms in that particular domain and logical statements that describe the relationships among the terms. WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet was created in the Cognitive Science Laboratory of Princeton university under the direction of psychology professor George Armitage Miller starting in 1985 and has been directed in recent years by Christiane Fellbaum. The project received funding from government agencies including the National Science Foundation, DARPA, the Disruptive Technology Office (formerly the Advanced Research and Development Activity), and REFLEX.

2.4 Homonyms:

A *homonym* is a word that is said or spelled the same way as another word but has a different *meaning*. For example
Sack (1) --a bag made of paper or plastic for holding customer's purchases
Sack (2) -- any of various light dry strong white wine from Spain and Canary Islands
In above given example both the „sack"s are spelled in the same manner but both have different meaning.
Similarly a single word can have just not only two but more than two meanings sometimes.

2.5 Word sense disambiguation:

WSD is a natural language processing problem. WSD removes ambiguity between multiple meanings of a word at a particular occurrence in a particular context. The word disambiguate divides into dis- not, ambiguous-unclear. Thus WSD identifies a particular sense for the particular occurrence of the word. In this first all of the different sense or meanings of the attribute are identified and then the correct meaning is extracted.

2.6 Indexing:

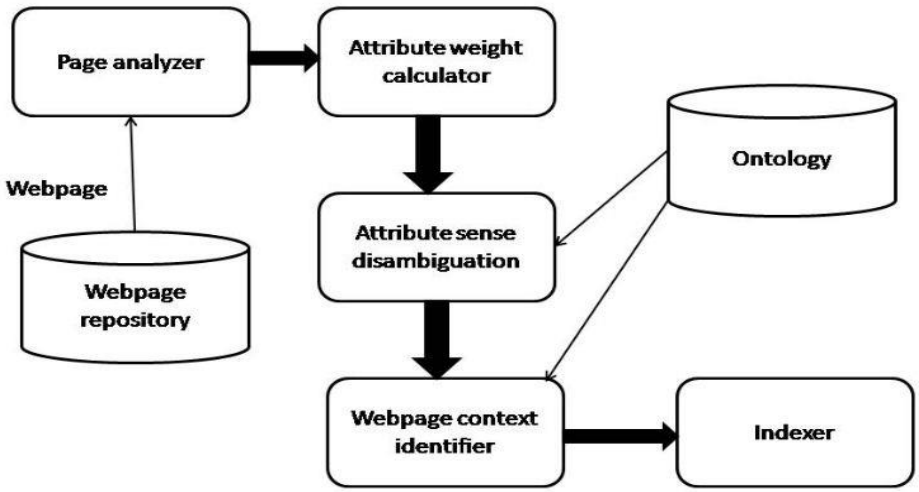
Indexing refers to create a list of data that can be quickly scanned. Indices save time as they make it faster to find a specific webpage and also serve as a guide to the webpages. The main advantage of storing an index is to optimize the speed and performance while finding relevant documents from the search engine storage area for a user given search query.

3. Overview and Architecture of the algorithm

3.1 Overview

In this proposed algorithm we are going to index the webpages alongside solving the problem of homonyms. The algorithm consists of five main modules- page analyzer, attribute weight calculator, attribute sense disambiguation, webpage context identifier and indexer. The control flows through these modules. Also the input is taken from a webpage repository and an ontological database is used by different modules for their respective tasks. A webpage is extracted from the repository and then goes through the algorithm making use of the ontology. Ontology is mainly used by two modules-
Attribute sense disambiguator and webpage content identifier.

3.2 Architecture



4. The proposed algorithm

Module-1 Page analyzer

In this step we analyze the webpage and then extract the attributes of the webpage by following the below given steps-
Extracting PIS (programmer intended search) attributes.PIS is extracted from internal identifiers, this gives what the programmer sees in the code of the webpage UIS (user intended search) attributes.
PIS is extracted from internal identifiers, this gives what the programmer sees in the code of the webpage. UIS is extracted from the options tag in html webpage code, this gives what the user sees on the html web document i.e. the free text on the webpage. Then using ontology the two extracted lists are intersected to get the final list i.e. FA (final attributes).These final attributes are the final result of this step and are served as input.

Module-2 Attribute weight calculation:

In the first step we get the attributes of the page which describe the relevance of that webpage for the domain we are working on. Now in this step we will calculate the term weight of each attribute. Term weight depends on the number of occurrences of the attribute in the page and also the number of occurrences of the synonyms and related words. Synonyms and related words are searched from ontology. Also even a single occurrence of the attribute is a hyperlink then it adds to the term weight and hence results in increasing its relevance.
The weight can be calculated in the following manner:

- Count the number of occurrences of the term W in the webpage. Let it be N.
- Count the number of occurrences of the synonyms and related words of W in the webpage. Let it be n.
- Create a set of words containing all the hyperlinks in the webpage say S. Then check if W or its synonym exists in the set. If it exists then set H as 1 else 0.
- Finally calculate the Term weight T.
  $T = N + 0.5n + H$

Now following the above steps we get term weight for each attribute. We will divide the attributes as primary and secondary on the basis of their term weights. The attribute with the highest term weight is the primary attribute and the rest are secondary attributes.

Module-3 Attribute Sense Disambiguation:

This step is purely based on word sense disambiguation. We disambiguate the meaning of all the attributes. In this step different sense or meanings or contexts of the attribute are extracted from ontology and then finally all the meanings of the attribute are listed together and then the correct meaning is identified.
LESK algorithm is used for identifying the correct meaning of the word. This algorithm is based on checking the neighborhood of the word and its meaning. Let our attribute be at some position x, then we check the meanings of all the terms that are present in the vicinity of the attribute at position x.. Now we calculate the amount of words that are in neighborhood of our attribute that match the correct context. Different combination synsets are made for getting the synset set, and one with the maximum overlap is selected.
Example:

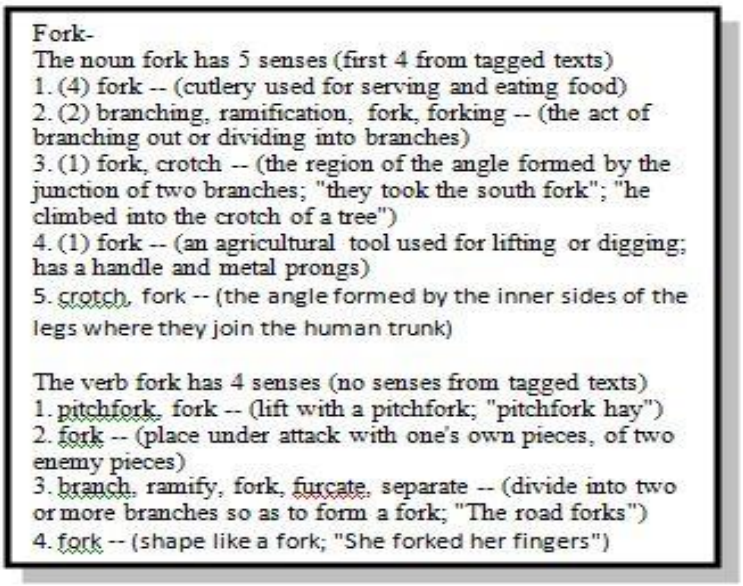


Figure 1 meanings of FORK as listed by Wordnet

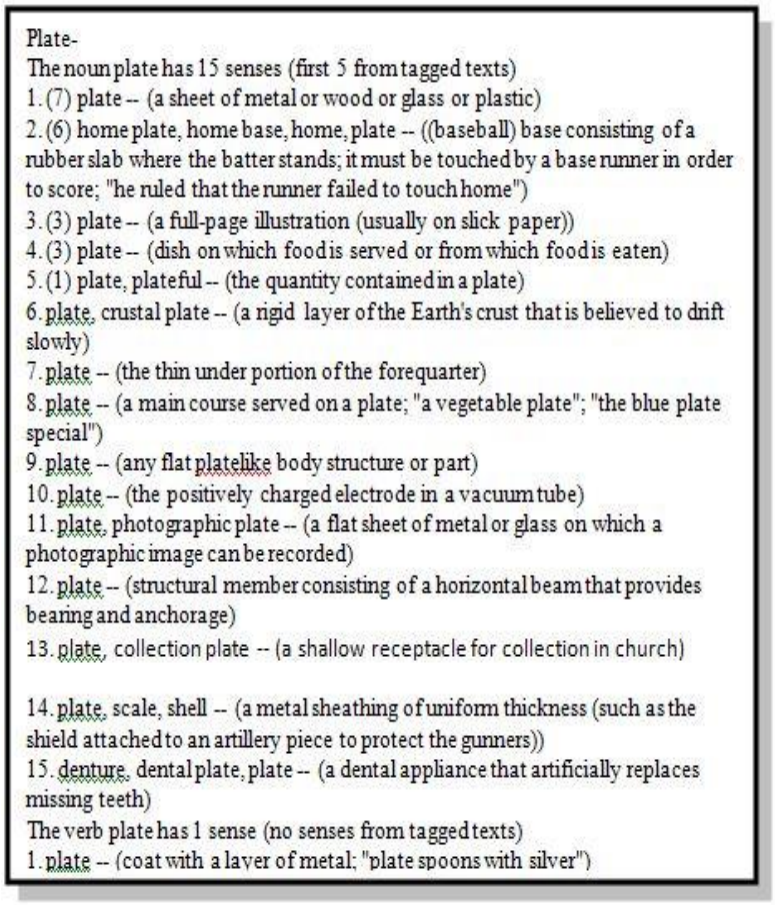


Plate-
The noun plate has 15 senses (first 5 from tagged texts)
1. (7) plate -- (a sheet of metal or wood or glass or plastic)
2. (6) home plate, home base, home, plate -- ((baseball) base consisting of a rubber slab where the batter stands; it must be touched by a base runner in order to score; "he ruled that the runner failed to touch home")
3. (3) plate -- (a full-page illustration (usually on slick paper))
4. (3) plate -- (dish on which food is served or from which food is eaten)
5. (1) plate, plateful -- (the quantity contained in a plate)
6. plate, crustal plate -- (a rigid layer of the Earth's crust that is believed to drift slowly)
7. plate -- (the thin under portion of the forequarter)
8. plate -- (a main course served on a plate; "a vegetable plate"; "the blue plate special")
9. plate -- (any flat platelike body structure or part)
10. plate -- (the positively charged electrode in a vacuum tube)
11. plate, photographic plate -- (a flat sheet of metal or glass on which a photographic image can be recorded)
12. plate -- (structural member consisting of a horizontal beam that provides bearing and anchorage)
13. plate, collection plate -- (a shallow receptacle for collection in church)
14. plate, scale, shell -- (a metal sheathing of uniform thickness (such as the shield attached to an artillery piece to protect the gunners))
15. denture, dental plate, plate -- (a dental appliance that artificially replaces missing teeth)
The verb plate has 1 sense (no senses from tagged texts)
1. plate -- (coat with a layer of metal; "plate spoons with silver")

Figure 2 Meanings for plate as listed by Wordnet

Fork#1 intersection with Plate#3 has the maximum similarity and overlap hence justifying the selection of the context if both these words are present in each other‟s neighborhood.
We use WordNet as the ontology and thus to extract different meanings we need to access the WordNet dictionary and for this purpose JWNL is used. JWNL- Java WordNet Library is a java API for accessing WordNet relational dictionary also provides morphological processing.

Module-4 Webpage context identifier:
In this step the catalog created in the previous step is analyzed and the meaning of the primary attribute is rechecked against the meaning of the secondary attributes of the same webpage which justifies the meaning identified in the above step and gives the correct context of the webpage.
For example:
Let our primary attribute is „jaguar‟ which can mean either jaguar animal or vehicle brand.
Then secondary attributes for

- Jaguar (animal) may be:- wildlife, animal, cat, etc.
- Jaguar (Luxury vehicle brand) may be:- automobile, cars, luxury brand, etc.

This justifies the sense of the attributes and also gives the correct context of the webpage i.e. in the example the page either relates to an animal belonging to the family of wild cats or it relates to vehicle brand for luxury cars.

Module-5 Indexer:
Finally after getting the correct meanings of the attributes and the context of the page the final resultant index is created that contains four columns.

- Column one is for word.
- Column two is for the context of the word.
- Column three tells if it is a primary attribute in those pages or secondary attribute.
- Last column is for the page ids containing that particular word with that particular meaning (context) and as primary/secondary attribute.

Output format:

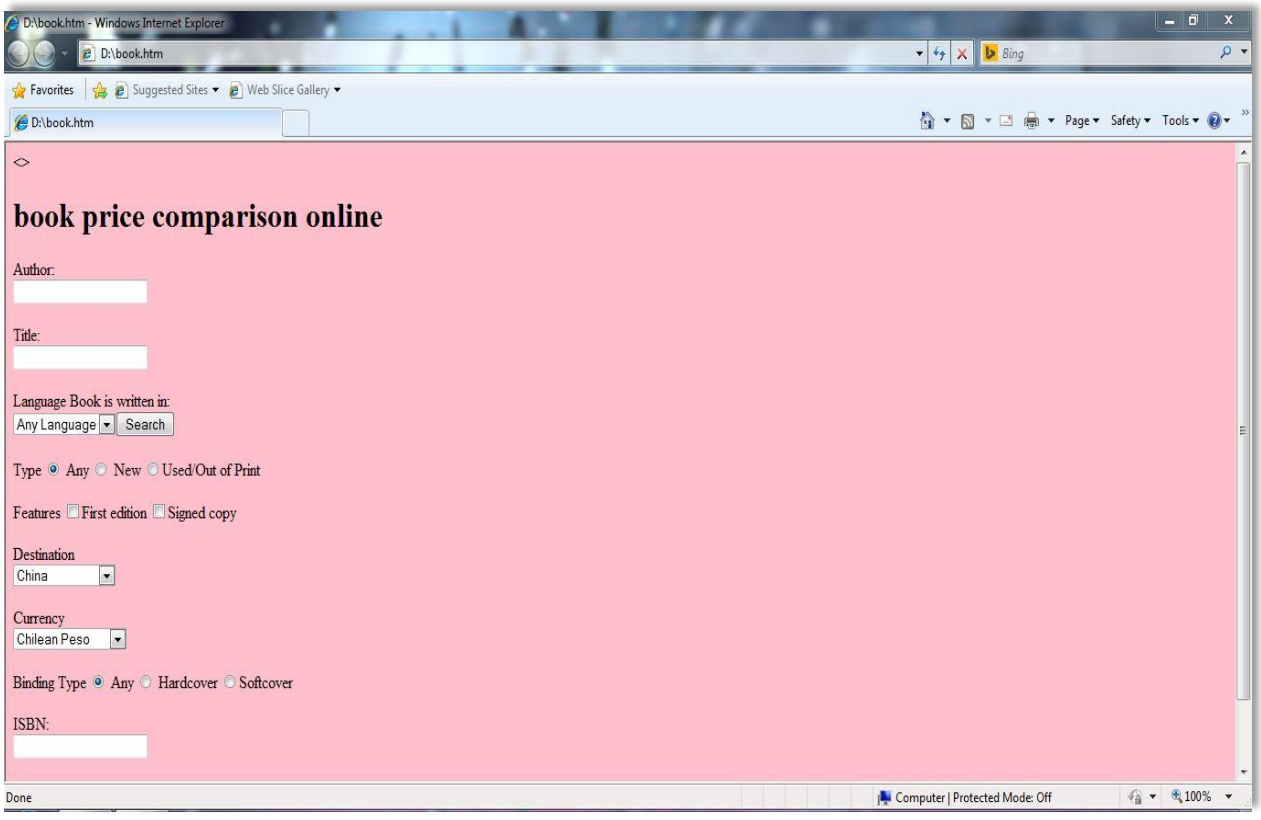
Word	Context	Primary/ Secondary	Page id

Figure 3 Output table

5. Experimental example

We have taken a prototype webpage form for this purpose.

Figure 4 Webpage prototype



This is the output table we get from the webpage form above-

Word	Context	Primary/ Secondary	Page id
Book	A publication, noun	Primary	1,2
Book	To arrange, verb	Secondary	1
Page	A sheet containing data, noun	Secondary	2
Author	Writer of a publication, noun	Secondary	1

Table-1 Output table for the above webpage

6. Conclusion and future work

The algorithm proposed indexes the webpage from hidden web using ontology. At first the page analyzer analyzes the webpage and extracts the attributes then for each attribute a numeric value is calculated on the basis of which we divide it into primary and secondary attribute. Then using word sense disambiguation all the possible meanings of the attributes are extracted so as to identify the correct one. Then to extract the context of the webpage from the primary attribute we recheck the sense of the secondary attributes. And finally an index is created that tells the context of the webpage too.

In the future, more semantics needs to be added to Deep Web processing, to achieve the goal of a Semantic Deep Web. For example, WordNet might be replaced or augmented by domain specific ontologies. While WordNet has excellent wide coverage and usability it lacks domain specific knowledge. We have reviewed existing ontologies for e-commerce and have not found any ontology with both the breadth and depth needed for this project. Thus we intend to build such an ontology in future work.

7. References

- AKSHR: A Novel Framework for a Domain-specific Hidden Web Crawler(Komal Kumar Bhatia, A.K. Sharma, Rosy Madaan Department of Computer Engineering, YMCA Institute of Engineering, Faridabad)
- Design of an Ontology Based Adaptive Crawler for Hidden Web (Manvi.M, K.K.Bhatia, A.Dixit)
- Hidden Web Data Extraction using Wordnet Ontology's(VidyaSagar Ponnam, V. P Krishna Anne, Venkata Kishore Konki)
- Lesk algorithm, Wikipedia
- Adapting the Lesk algorithm for word sense disambiguation to Wordnet(Satanjeev banerjee, university of Minnesota, USA)
- Automatic Attribute Extraction from deep web data sources(Yoo Jung An, James Geller, Yi Ta Wu, Soon Ae Chun)
- Webpage Indexing based on Prioritized Ontology Terms(Sukanta Sinha, Rana Dattagupta, Debajyoti Mukherjee, TCS, Kolkata, India)
- Conception and use of Ontologies for indexing and searching by Semantic content of video courses(Merzougui Ghalia, Djoudi Mahieddine, Behaz Amel, Batna University, Algeria)

