Extraction of Product Information for Trustworthiness

Pooja Vanerkar, Prof. Chaita Jani

[1]Computer Engineering, Kalol Institute of Technology, Gujarat Technological University

[2]Professor, Kalol Institute of Technology, Gujarat Technological University

Abstract: ***Online market is helpful to the consumers and brings wealth of information. However, online market is also flooded with fake and shoddy products. Trustworthiness is a solution to solve these problems which is raised by the authors' laboratory. The model of trustworthiness evaluates the product in three dimensions, which are authenticity, reliability and credibility. But the data to analyze trustworthiness evaluation is mainly from manual import. Product Information Object is the organic whole of product information and product reviews. In this paper we simulate and show a compression between algorithms which can show the discrimination accordingly.***

Keywords: Trustworthiness, Product extraction, accuracy, framework, product information

1. Introduction

Web search engines have become the most convenient way to help people discover their desired contents in this huge collection of information. Searching for product information on web search engine, however, is clearly inadequate and has some limitations. Search results returned from the web search engine only match the similarity between query terms and content of web pages. A business entity will be able to search the suppliers offer space and to filter the ones that are fitted to the entity current needs, in a scale of the size of the Internet. The growth in the number of users who routinely use the web to search for information on goods and services continues unabated. In this expanded use of the web it is not uncommon for end users to view the web as a database and search for information in ways that cannot be directly accomplished by traditional keyword-based web search engines.

2. Literature Review

A. ***Title:*** *An Algorithm of Product Information Extraction from Web Pages: a Document Object Model Analysis Approach*

Algorithm: Extraction Based on Sub-Tree (exbst) and Extraction Based on Sibling Nodes (exbsn).

Problem description: One of the main problems that vertical search systems and entity search systems encounter with is how to extract useful information that helps them to perform more efficiently and more effectively. For this particular study, the goal of product search engines is to present useful information about products for their user. Thus, the product search engines need to accurately identify product information web pages and extract important information from these pages. The important information includes product name, product description, product image, and product price.

Advantages: To access the performance of the algorithm there accuracy is examined, the accuracy indicates the effectiveness of the algorithm. The combination of exbst and exbsn algorithm provides a higher accuracy in product information extraction from web page and outperforms when each of them extracts the information by its own.

Disadvantages: It causes an error in the combined algorithms from the performance of the *is productimage(imgobj)* function and some authoring style of web pages such as presenting product information using iframe tag and a complex table can result in errors in extracting a product information object. There is less robustness in extraction feature based on DOM tree structure.

B. ***Title:*** *Extracting accurate data from multiple conflicting information on web sources*

Algorithm: Veracity problem is formulated which discover true fact for the information and framework to solve

Problem description: The problem of trustworthiness these are formulated by truth finder algorithm.To design a system this finds true facts among conflicting information, and identifies Trust worthy websites better than the popular websites. In this we assign ratings based on two things- popularity or the hits & number of occurrences of same data. As we can't give preference only to popularity, we have considered another rating i.e. about number of occurrences of same data in several other websites which are less popular.

Advantages: Here Fact Finder achieves very high accuracy in discovering true facts. It can select better trustworthy websites than authority-based search engines such as Google.

C. ***Title:*** *A Secure Composition Framework for Trustworthy Personal Information Assistants*

Algorithm: In this a framework is used that supports composition of individual agents system such as Personal Information Assistants (PIA) that enables users to accomplish complex tasks for this it uses winagent system.

Advantages: Benefit of this approach is that in the framework the personal information handled by the agent system is guaranteed to be free from accidental leakage to websites that are not trustworthy, thereby ensuring the privacy of end-user data. Some other benefits are: Enables

compensability and re-use, Eases creation and deployment, Untrusted site blocker, Sensitive data tracker.
Disadvantages: There is a security issue as the pias are highly data driven

D. ***Title:*** *A Survey on Unsupervised Extraction of Product Information from Semi-Structured Sources*
Algorithm: The algorithm is based on a clustering approach that uses Structural and visual features of web page elements. Wrapper generation system such as roadrunner, SG-WRAP, X-WRAP, dela.
Advantages: All the techniques of Wrapper generation system are very efficient, the visual representation provides the most valuable clustering features.
Disadvantages: Information extraction from cross – website becomes more complex when it moves towards semantic web.

E. ***Title:*** *Extraction of Product Information Object for Trustworthiness*
Algorithm: Extraction algorithm of product Reviews base on hidden web
Advantages: The algorithm uses the feature of location of product reviews and the browsing path of hidden product reviews to achieve automated extraction of hidden product reviews, and its precision is relatively high.

3. Proposed Work

- Convert the URL of the product or the keyword into the source code so that we can extract the product information.
- Convert the HTML page into Dom tree, the leaf nodes in the DOM tree contain text content and format information, this also removes java script so the complexity also reduces.
- Match the product semantic dictionary with the OWL FULL, if the semantic matches they are forwarded further for checking other criteria, if it doesn’t matches the product is not trustworthy.
- Calculate the frequency of the key phrase with the help of Td-idf, key phrase with higher degree are taken into consideration and other products are neglected.
- As the degree is calculated rating is done for the key phrase the products and the flags are set
- Check the common attributes such as phone no., email id, address etc but it’s not compulsory so the rating of the product will be described according and the percentage will be displayed accordingly.

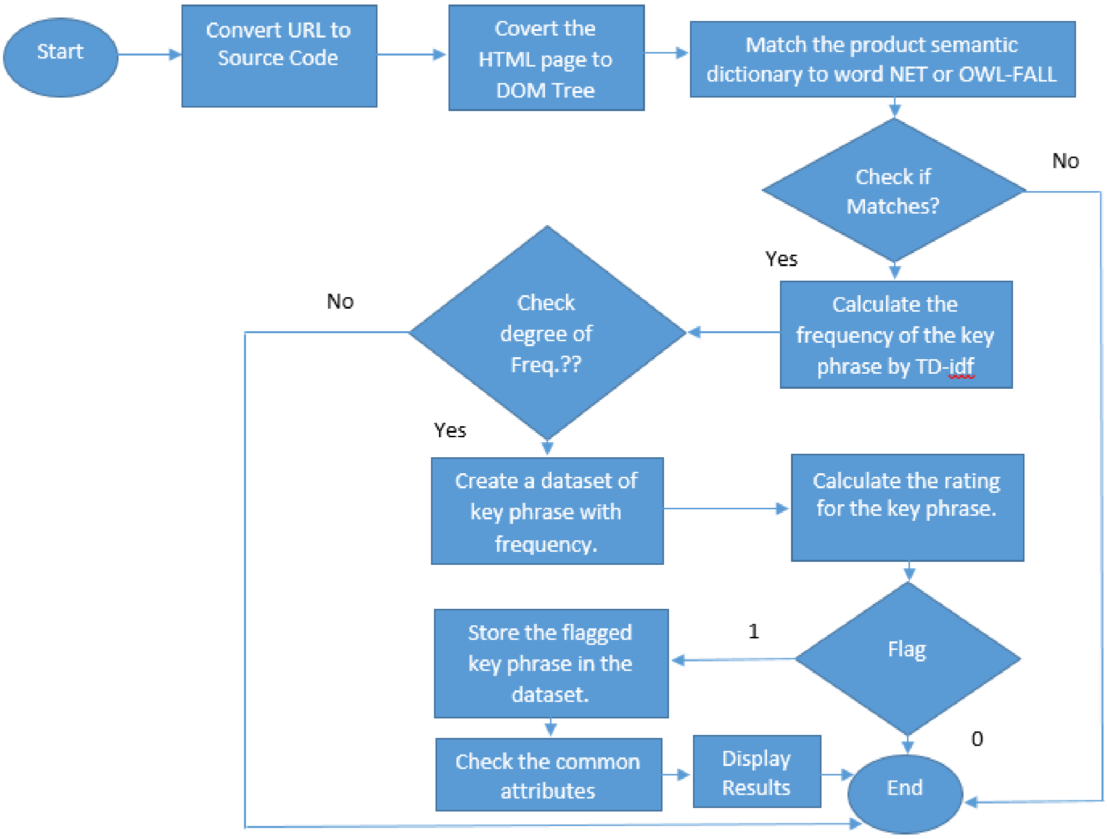
Start
Convert URL to Source Code
Covert the HTML page to DOM Tree
Match the product semantic dictionary to word NET or OWL-FALL
Check if Matches?
No
Yes
Calculate the frequency of the key phrase by TD-idf
Check degree of Freq.??
No
Yes
Create a dataset of key phrase with frequency.
Calculate the rating for the key phrase.
Flag
1
0
Store the flagged key phrase in the dataset.
Check the common attributes
Display Results
End

Figure 1: Proposed flowchart

4. Conclusion and Future Work

- The main objective of this experiment is to combine multiple techniques and achieve accurate result of trustworthiness is the baseline of this study.
- Checking and justifying that the product is not trustworthy on the bases of common attributes of the product is not proper, so here an average of rating, frequency of degree, checking the product semantic dictionary and common attributes is taken into consideration the accuracy of the result increases.
- In this the flow is presented as the back end and its shows the path and strategy how to use this method by using appropriate front end to get efficient result.

References

[1] **An Algorithm of Product Information Extraction from Web Pages:a Document Object Model Analysis Approach** Worasit Choochaiwattana□ □ □*2012 2nd International Conference on Information Communication and Management (ICICM 2012)**IPCSIT vol. 55 (2012) © (2012) IACSIT Press, Singapore** DOI: 10.7763/IPCSIT.2012.V55.19*

[2] **Extracting accurate data from multiple conflicting Information on web sources Akshata angadi1, karuna gull2, padmashri desai 3** 1,3computer science and engineering department, 1k.l.e.i.t. , #3 b.v.b.c.e.t. hubli, india 2k.l.e.i.t. hubli, india E-mail: 1akshata_angadi@yahoo.co.in, 3padmashri@bvb.edu, 2karuna7674@gmail.com

[3] **A Secure Composition Framework for Trustworthy Personal Information Assistants_**V.N. Venkatakrishnan Wei Xu I.V. Ramakrishnan R. Sekar Department of Computer Science Department of Computer Science University of Illinois at Chicago Stony Brook University venkat@cs.uic.edu fweixu, ram, sekarg@cs.sunysb.edu

[4] A Survey on Unsupervised Extraction of Product Information from Semi-Structured Sources Abhilasha Bhagat , Vanita Raut *ME Computer Engineering, Assistant Professor Dept. of Computer Engineering G.H.R.I.E.T., Savitribai Phule University, pune G.H.R.I.E.T., Savitribai Phule University, pune PUNE , India PUNE , India*

[5] Extraction of Product Information Object for Trustworthiness Shenglong Mi National Engineering Lab for Ecommerce Technologies Fudan University, Shanghai, China mishenglong2008@gmail.com Yinsheng Li*, Hao Chen, Yong Fang National Engineering Lab for Ecommerce Technologies Fudan University, Shanghai, China liys@fudan.edu.cn.

[6] Securing Web Application Code through Static Analysis and Runtime Protection. Yao-wen huang and fang yu and christian hang and chung-hung tsai and der-tsai lee and sy-yen kuo. In *Thirteenth World Wide Web Conference, New* York City, 2004.

[7] V.N. Venkatakrishnan. *Enforcement Techniques for Expressive Security Policies.* PhD thesis, Stony Brook University, 2004.

[8] D. Volpano, G. Smith, and C. Irvine. A sound type system for secure flow analysis. *Journal of Computer Security (JCS), 4(3):167–187, 1996.*