IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. IX (May-Jun. 2014), PP 54-60
www.iosrjournals.org

Content Evocation Using Web Scraping and Semantic Illustration

Vasani Krunal A.
(Department Of Computer Engineering, R.K. University, Rajkot India)

Abstract: *Web scraping is the process of automatically collecting information from the World Wide Web. It is a field with active developments, sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text processing, artificial intelligence and human-computer interactions. It means extraction of content from different web pages using web scrapping and semantic illustration. Web Scrapping is a process of evocation of content from HTML pages and related to web indexing. A commonly used measure for tree similarity is the tree edit distance which easily can be extended to be a measure of how well a pattern can be matched in a tree. An obstacle for this approach is its time complexity, so we consider if faster algorithms for constrained tree edit distances are usable for web scraping, and to reduce the size of the tree representing the web page. Different applications of web scraping are used by current market to achieve best web scraping output, Like Web Data Extraction, Data Collection, Screen Scraping. Many different algorithms are used for web scraping like "tree pattern matching", "tree mapping", "approximate tree matching". But in general "tree edit distance" algorithm is used. But with this algorithm many issues of incorrectness of data, low efficiency and higher time complexity have analyzed. In this research I am focus to improve the performance of tree edit distance problem. And I am also trying to focus on higher bound time complexity of this algorithm.*

Keywords: *web scraping; data mining; tree edit distance*

I. Introduction

Web scraping is the process of automatically collecting information from the World Wide Web. It is a field with active developments sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text processing, semantic understanding, artificial intelligence and human-computer interactions.

The pages being scraped may embrace metadata or semantic markups and annotations, which can be used to locate specific data snippets. If the annotations are embedded in the pages, as Micro-format does, this technique can be viewed as a special case of DOM parsing. In another case, the annotations, organized into a semantic layer, are stored and managed separately from the web pages, so the scrapers can retrieve data schema and instructions from this layer before scraping the pages.

DOM parsing: By embedding a full-fledged web browser, such as the Internet Explorer or the Mozilla browser control, programs can retrieve the dynamic contents generated by client side scripts. These browser controls also parse web pages into a DOM tree, based on which programs can retrieve parts of the pages.

Web Indexing – Search engines such as Google and Bing routinely scour the internet collecting data on Web sites to ensure that the most relevant match is retrieved in response to your search. They use a combination of mathematics, linguistics and psychology to determine what information available is the most relevant to your post. Most search engines will re-index their results about once a month to stay up to date.

Googlebot is Google's web crawling robot, which finds and retrieves pages on the web and hands them off to the Google indexer. It's easy to imagine Googlebot as a little spider scurrying across the strands of cyberspace, but in reality Googlebot doesn't traverse the web at all. It functions much like your web browser, by sending a request to a web server for a web page, downloading the entire page, then handing it off to Google's indexer. Googlebot gives the indexer the full text of the pages it finds. These pages are stored in Google's index database. This index is sorted alphabetically by search term, with each index entry storing a list of documents in which the term appears and the location within the text where it occurs. This data structure allows rapid access to documents that contain user query terms.

To improve search performance, Google ignores (doesn't index) common words called *stop words* (such as *the*, *is*, *on*, *or*, *of*, *how*, *why*, as well as certain single digits and single letters). Stop words are so common that they do little to narrow a search, and therefore they can safely be discarded. The indexer also ignores some punctuation and multiple spaces, as well as converting all letters to lowercase, to improve Google's performance.

Tree Pattern Matching Algorithm Working for web scrapping

I investigate the potential of using approximate tree pattern matching based on the tree edit distance and constrained derivatives for web scraping. I argue that algorithms for constrained tree edit distances are not suited for web scraping. To address the high time complexity of optimal tree edit distance algorithms, The lower bound pruning algorithm which, based on the data tree T_D and the pattern tree T_P , will attempt to remove branches of T_D that are not part of an optimal mapping. Its running time is $O(| T_D || T_P | \cdot \sigma (T_D, T_P))$, where $\sigma (T_D,T_P)$ is the running time of the lower bound method used. Although it asymptotically is close to the approximate tree pattern matching algorithms, In practice the total execution time is reduced in some cases. Some further development is require for several methods for determining a lower bound on the tree edit distance used for approximate tree pattern matching.

Challenge For Web Scrapping By Improving Tree Pattern Matching Algorithm.

The data extraction experiment showed that there are cases where the approximate tree pattern matching algorithm extracts incorrect data. To avoid this and thus cover more cases, a final solution requires more work. So My research work is to find out better methods for finding out lower bounds for the tree edit distance with cuts (Minimum or Maximum Cuts of Tree) which will also be helpful to extract efficient and correct data.

1.1 main area of web scraping

There are two primary channels for distributing ads: sponsored search (or paid search advertising) and contextual advertising (or content match). Sponsored search advertising displays ads on the page returned from a Web search engine following a query; whereas contextual advertising displays ads within the content of a generic, third part, Web page. A commercial intermediary, namely ad network, is usually in charge of optimizing the selection of ads with the twofold goal of increasing revenue and improving user experience. The ads are selected and served by automated systems based on the content displayed to the user.

1.2 brief about web scraping

Usually, Web scrapers simulate human exploration of the World Wide Web by either implementing low-level hypertext transfer protocol or embedding suitable Web browsers. Web scraping is closely related to Web indexing, which is an information retrieval technique adopted by several search engines to index information on the Web through a bot. In contrast, Web scraping focuses on the transformation of unstructured data on the Web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet.

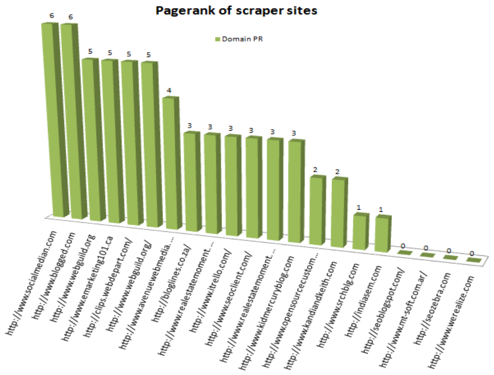


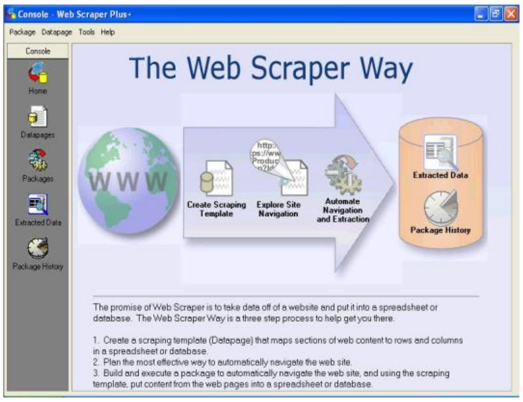
Fig. 1 Page Rank Of Scraper Sites

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

II. Problem definition

A program that detects same behavioural templates in a particular information source, extracts its content and translates it into a relational form is called a wrapper. Currently web scraping is used for online price comparison, weather data monitoring, website change detection, Web research, Web mashup, and Web data integration. Web scraping specially use for information extraction. A form of information extraction is text mining, an information retrieval task aimed at discovering new, previously unknown information, by automatically extracting it from different text resources.

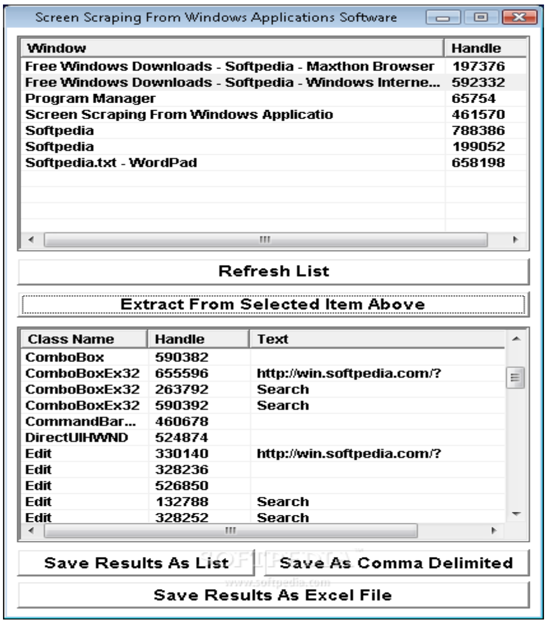
Screen Shot Of Web Scraping Process



The promise of Web Scraper is to take data off of a website and put it into a spreadsheet or database. The Web Scraper Way is a three step process to help get you there.

1. Create a scraping template (Datapage) that maps sections of web content to rows and columns in a spreadsheet or database.
2. Plan the most effective way to automatically navigate the web site.
3. Build and execute a package to automatically navigate the web site, and using the scraping template, put content from the web pages into a spreadsheet or database.

Fig. 2 Web Scraping Console


Screen Scraping From Windows Applications Software

Window	Handle
Free Windows Downloads - Softpedia - Maxthon Browser	197376
Free Windows Downloads - Softpedia - Windows Interne...	592332
Program Manager	65754
Screen Scraping From Windows Applicatio	461570
Softpedia	788386
Softpedia	199052
Softpedia.txt - WordPad	658198

Refresh List
Extract From Selected Item Above

Class Name	Handle	Text
ComboBox	590382	
ComboBoxEx32	655596	http://win.softpedia.com/?
ComboBoxEx32	263792	Search
ComboBoxEx32	590392	Search
CommandBar...	460678	
DirectUIHWND	524874	
Edit	330140	http://win.softpedia.com/?
Edit	328236	
Edit	526850	
Edit	132788	Search
Edit	328252	Search

Fig. 3 Screen Scraping from windows application

2.1 Algorithm

Main algorithm of particular research area is “Tree edit distance algorithm”. It also shows that how minimizes the total number of forest which are used in Zhang-Shasha and Klein strategies. Further development is require for several methods for tree edit distance which gives structured and efficient data from the given searched string.

III. Literature Survey

Journal of Advanced Computer Research presented at December-2012. This journal is about two areas, first semantic web and web mining to improve mining and using mining creating semantic. It list outs the benefits, challenges and opportunities in area of web mining. It also represents the idea to improvise the results of web mining by taking advantage of the new semantic structures on the web; and also making use of web mining, for building up the semantic web by extracting similar meaning, useful patterns, structures and semantic relation from existing web resources. It gives challenge that “The critical aspect in today’s business scenario is how data is converted into Information and subsequently how information is converted into knowledge”. [1]

International journal of research in engineering and applied science presented at February-2012. It defines the exact meaning of web scraping. Web scraping is a process of extracting useful information from HTML pages. Web scraping is implemented with powerful string matching operations. This literature mainly gives knowledge of different techniques of web scraping.

- Scraping can be performed for UNIX.
- Prolog Server Pages

This paper represents the technique to scrap HTML pages and utilize it as per requirement and as per data type.This paper requires high performance testing so it's cost becomes very high. [2]

The paper from www.sciedu.ca/air Artificial Intelligence Research in the year of 2013. It is about to explain web advertising, Collaborative filtering and web scraping. According to this paper, Web scrapers simulate human exploration of the World Wide Web by either implementing low-level hypertext transfer protocol or embedding suitable Web browsers. Web scraping is closely related to web indexing, which is an information retrieval technique adopted by several search engine to index information on the web through a bot. Web scraping focuses on the transformation of unstructured data on the Web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. [3]

The main algorithm of particular research area "Tree edit distance algorithm". It also shows that how minimizes the total number of forest which are used in Zhang-Shasha and Klein strategies. Paper from artificial intelligence research which I have mentioned as a third literature also demonstrate technical steps of web scraping programming.

- Using that steps we can analyze the flow of web scraping.
- Following are the steps:
- First step is to read a page from a particular web site.
- Second step is table lookup which finds the proper words from title and analyze that words.

There are many alternative table lookup techniques available, including hash tables and data step MERGE or PROC SQL.

- Third step will split words from a title.
- Forth step is for counting and categorizing.

There are so many words with very low frequencies, we arbitrarily decide to process only those words with three or more occurrences. This is easily done by applying the WHERE= DATASET OPTIONS to the OUT= dataset in the TABLES statement.

In this way it identifies all type of words with different frequencies. The data step above implements scaling. This manipulation requires two passes over the input data. The first pass finds the maximum frequency; the second scales using the found maximum. In the data step, we use two DO UNTIL loops with a SET statement inside each.

- Finally It will give HTML output.
- Limitation:

Extracting keywords given text (also known as, term extraction or keyword extraction) is an actively developed and commercially promising research area of data mining. Various innovative algorithms exist and some demonstrations are available free of charge on the web, at the time of writing this paper.

Web scraping is a process of extracting useful information from HTML pages. Web scraping is implemented with powerful string matching operations. This literature mainly gives knowledge of different techniques of web scraping.

- Scraping can be performed for UNIX.

Prolog Server Pages.[5]

Algorithm "Tree edit distance algorithm" also shows that how minimizes the total number of forest which are used in Zhang-Shasha and Klein strategies.

Further development is require for several methods for tree edit distance which gives structured and efficient data from the given searched string. [4]

Robust Algorithm for the Tree Edit Distance

It achieves both important concepts optimal worst-case complexity Efficient. It computes the minimum-cost sequence of node edit operations that transform one tree into another. [6]

K.C Tai's paper presents the first algorithm to solve the tree edit distance problem. But it is complicated and impractical to implement.

- Its time complexity is:
- $O(|T_1||T_2|.\ \text{height}\ (T_1)^2\ .\ \text{height}(T_2)^2)$

Which in the worst case is $O(|T_1|^3\ |T_2|^3)$. [7]

3.1 Open Issues

In web scraping any developer wants to extract data from different sources like from different web sites. Most of developer uses regexes means regular expression. PHP also gives some functions like explode.

Following are number of techniques that gives number of parsing and scripting code that can perform operation of data extraction.

Technique no. 1:Text searching (no regexes)

This technique works on trial and error technique. Code of Parsing :

```
text = ""
File.open("body_serialized.ser") do |file|
        text = Marshal.load(file)
end
text.each_line do |line|
        line = line.downcase
        if(line.index("title=\"view all posts"))
                contents = line.split(" ")
                interesting = contents[9]
                puts interesting
end         end
```

Running the script produces the following output :

```
        <a
        nil
        http">http</a>
        java">java</a>
        music">music</a>
        ruby">ruby</a>
        spider">spider</a>
        swing">swing</a>
        synchronization">synchronization</a>
        threads">threads</a>
                  videos">videos</a>
```

This output gave values but with some HTML tags that we do not want. So to remove these unnecessary tags we can add some more script in above and using that can get following output.

```
        http      java
        music     ruby
        spider    swing
        synchronization   threads
          videos
```

Advantages of this technique:

No need to understand regexes

Do not need any third-party library/framework/tool

It’s easy to do if we know a little of programming.

Disadvantages of this technique :

Searching text in this way takes a lot of tries to get to the right results

If the site’s html changes it’s structure , the script is unusable (you would have to do this all over again)…(That’s the big issue in the field of web scraping.)

Technique no. 2 : Text searching (with regexes)

If we have a knowledge of regexes, this technique will be a lot simpler & faster . Here the script that extracts the categories using regular expressions :

```
text = ""
File.open("body_serialized.ser") do |file|
        text = Marshal.load(file)
end
matches = text.scan(/title=\"View all posts.*?>(.*?)<\/a>/i)
matches.each do |match|
        puts match          end
```

Advantages of this technique

Write less code

Development speed increases. This script takes about 2 minutes to write (and it worked the first time) , while the script for the first technique took about 10 minutes.

The regex is pretty easy to replace (If site will change it's HTML structure then no need to replace whole script just we have to change regexes.)

Disadvantages of this technique

It needs the knowledge of regexes to use this technique.
A lot of developers don't know how to use them.

Technique no. 3 : scraping using CSS (kind of)

We could scrape a web page "with style". It means we could use the CSS selectors to find the information we need . We can find more informations about it.

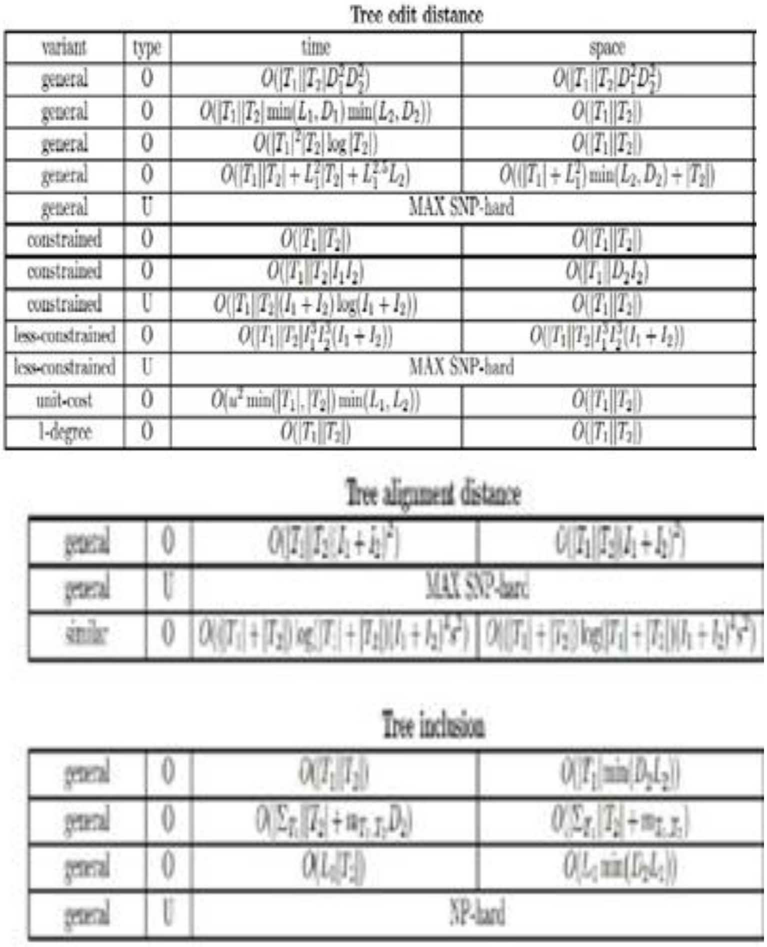
Advantages of this tecnique

If we know how to use CSS , development speed would increase.
A page's style doesn't change but HTML structure changes often.

Disadvantages of this technique

We must know CSS.

3.2 Complexity Comparison Of Various Tree Edit Distance



Tree edit distance

variant	type	time	space
general	0	$O(\|T_1\|\|T_2\|D_1^2D_2^2)$	$O(\|T_1\|\|T_2\|D_1^2D_2^2)$
general	0	$O(\|T_1\|\|T_2\|\min(L_1,D_1)\min(L_2,D_2))$	$O(\|T_1\|\|T_2\|)$
general	0	$O(\|T_1\|^2\|T_2\|\log\|T_2\|)$	$O(\|T_1\|\|T_2\|)$
general	0	$O(\|T_1\|\|T_2\|+L_1^2\|T_2\|+L_1^{2.5}L_2)$	$O((\|T_1\|+L_1^2)\min(L_2,D_2)+\|T_2\|)$
general	U	MAX SNP-hard	
constrained	0	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\|T_2\|)$
constrained	0	$O(\|T_1\|\|T_2\|I_1I_2)$	$O(\|T_1\|D_2I_2)$
constrained	U	$O(\|T_1\|\|T_2\|(I_1+I_2)\log(I_1+I_2))$	$O(\|T_1\|\|T_2\|)$
less-constrained	0	$O(\|T_1\|\|T_2\|I_1^3I_2^3(I_1+I_2))$	$O(\|T_1\|\|T_2\|I_1^3I_2^3(I_1+I_2))$
less-constrained	U	MAX SNP-hard	
unit-cost	0	$O(u^2\min(\|T_1\|,\|T_2\|)\min(L_1,L_2))$	$O(\|T_1\|\|T_2\|)$
1-degree	0	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\|T_2\|)$

Tree alignment distance

general	0	$O(\|T_1\|\|T_2\|(I_1+I_2)^2)$	$O(\|T_1\|\|T_2\|(I_1+I_2)^2)$
general	U	MAX SNP-hard	
similar	0	$O((\|T_1\|+\|T_2\|)\log(\|T_1\|+\|T_2\|)(I_1+I_2)^4s^2)$	$O((\|T_1\|+\|T_2\|)\log(\|T_1\|+\|T_2\|)(I_1+I_2)^4s^2)$

Tree inclusion

general	0	$O(\|T_1\|\|T_2\|)$	$O(\|T_1\|\min(D_2L_2))$
general	0	$O(\Sigma_{T_1}\|T_2\|+m_{T_1,T_2}D_2)$	$O(\Sigma_{T_1}\|T_2\|+m_{T_1,T_2})$
general	0	$O(L_1\|T_2\|)$	$O(L_1\min(D_2L_2))$
general	U	NP-hard	

Fig 6 Time Complexity Of Different Tree Edit Distance

3.3 Existing Algorithm Steps

When the node is encountered, so the least-cost mapping can be selected right away. To accomplish this the algorithm keeps track of the keyroots of the tree, which are defined as a set that contains the root of the tree plus all nodes which have a left sibling.

The keyroots of a tree are determined in advance, allowing the algorithm to separate the concepts of tree distance and forest distance.

Tree distance - The distance between two nodes when considered in the context of their left siblings in the trees T and T′.

Forest distance - The distance between two nodes considered separately from their siblings and ancestors but not from their descendants.

For each node, the calculation to determine the minimum cost mapping from the node to a node in the other tree (the tree distance) depends only on mapping the nodes and their children. To determine the minimum cost mapping of a node, then, one must know the minimum cost mapping from all the keyroots amongst its children, plus the cost of its leftmost child (the forest distance of its rightmost child). Since the nodes are numbered in postorder, the algorithm proceeds in the following steps:
1. Determine the mappings from all leaf keyroots.
2. Recursively determine the mappings for all keyroot at the next higher level.
3. Determine the mapping of the root.
The output of the algorithm is the sequence of edit operations that transforms the first tree into the other with minimal cost.

3.3 Logical Steps Of Proposed Solution

As mentioned above main three operations are required for tree edit distance. In that three steps because of recursive mapping functions it comes to upper-bound complexity. So that recursive function can be minimized upto the required level or it can be suspended when minimal mapping cost will be found.

1. As shown in pseudo code first compute leftmost leaf descendant of the subtree with all keyroots.
2. Upto all the keyroots find FD (forest distance) of nodes.
3. From all possible FD compute the minimum FD.
4. In existing algorithm it applies to all the keyroots. Which can be modified, Mapping function can be applied on required node of tree for the specific operarttion with consideration of cost.
5. If minimal mapping cost is found and if it is applicable then there is no need to apply mapping on all the keyroots.
6. Determine lower bound of all selected minimal cost mapping.(Currently it is only working on higher bound, that I am)

Determine the mapping for required keyroots upto next level as per given in the searched string. (using any lower bound method that I am still surveying). with a ratio of quantities and units.

IV. Conclusion

After analyzing all these methods and techniques of web scraping, I conclude that “Tree edit distance” algorithm extremely helpful to find out lower bound technique and also to improve performance of extracted data from web. As well as using these technique I can analyze all the operations regarding web scraping which will be useful to develop efficient method for web scraping.

Acknowledgement

The accomplishment of this paper work is not only due to my efforts. In fact, I stand on the shoulders of many people. So, with great pleasure I take this opportunity to express my gratitude towards all the individuals who have helped and inspired me in my work.

I am deeply indebted & would like to express my gratitude to my research guide **Prof. Kamlesh Patel, From Computer Engineering Department Of R.K. University, Rajkot,** for his great efforts and instructive comments in this work. He has devoted significant amount of his valuable time to plan and discuss the dissertation work. During the dissertation he provided me excellent guidance & support during discussion about my progress. He allowed me a great deal of freedom to choose a research topic, focusing more on my research interests and building skills that will allow me to be a successful in my life. Without his experience and insights, it would have been very difficult to do quality work.

References

[1]. Sanjay Kumar Malik, Sam Razvi: Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation, International Conference on Computational Intelligence and Communication Systems, 2011.
[2]. Syeda Farha Shazmeen, Etyala Ramyasree: Semantic Web Mining: Benefits, Challenges and Opportunities, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-7 December-2012.
[3]. Parminder Pal Sing Bedi, Sumit Kumar: Web scraping and implementation using prolog server pages in semantic web, IJREAS Volume 2, Issue 2 ISSN: 2249- 3905, February-2012.
[4]. Eloisa Vargiu, Mirko Urru: Exploiting web scraping in a collaborative filtering based approach to web advertising, www.sciedu.calair, Artificial Intelligence Research, Vol 2, No. 1, 2013, online published at December 5,2012.
[5]. Sareh Aghaei, Mohammad Ali Nematbakhsh, Hadi Khosravi Farsani, International Journal of Web & Semantic Technology(IJWest) Vol 3, No.1, January 2012.
[6]. K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems" SIAM Journal of Computing, Vol 18-6, (1989), p. 1245-1262.
[7]. Mateusz Pawlik, Nikolaus Augsten, Proceedings of the VLDB Endowment (PVLDB), Vol. 5, No. 4, pp. 334-345 (2011),