Social Network Wrappers (SNWs): An Approach used for Exploiting and Mining Social Media Platforms

Mamta Madan, Ph.D
Professor
Vivekananda Institute of
Professional Studies
GGSIP University

Meenu Chopra
Assistant Professor
Vivekananda Institute of
Professional Studies
GGSIP University

ABSTRACT
This paper tries to portrait outline study on the detailed approaches which are related to the working of a Social Media Networks Extraction System (SMNES) or the Social media (SM) platform with the perception of Social Network Wrappers (SNWs) and their issues like creation, perpetuation and support etc. In this paper we discuss in detail the obstacle related to the generation or creation of SNWs, initiation and support, and other important approaches. At last, we discuss the problem related to SNWs maintenance; propose our recommendation in adapting Social Network Wrappers fully automatically (SNWs).

Keywords
Social Media Networks (SMNs), Social Media Network Extraction System (SMNES), Social Network Wrappers (SNW)

1. EARLIER APPROACHES TO SOCIAL MEDIA NETWORKS (SMN)
The various approaches have been exploited by many researchers, but the first one to extract the data from SMNs was from Information Extraction (IE) approaches. Sarawagi [1] calls them rule-based or analytical approach and human-coded or learning-based approach respectively. These descriptions define the similar notion or the idea: the first method, particularly, the rule-based ones, is used to develop a system in which a strong closeness with both the requisites and the services is needed, so the human presence is crucial. Analytical methods are more efficient and dependable in domains of unorganized structure (like natural language processing problems, facts extraction from speeches, and automated text categorization [2]). Kaiser and Miksch [3] segregate them into two categories, firstly, knowledge engineering approaches and Secondly, Informational approaches.
Also in some approaches of the latter family, it is used to develop a system that requires end-user proficient to define rules (usually program snippets or regular expressions) to perform the extraction. In learning-based as well as in human-coded approaches areas of expertise are needed: people defining rules and practicing the system must have experience in programming and good domain knowledge.

2. SOCIAL NETWORK WRAPPERS (SNWS)
The SNWs is a process or system that accomplishes a family of algorithms, which hunt and discover the end-user expected information that needs to obtain from an unorganized Social Media Network (SMN), and convert them into organized data, *blending and binding* this extract information for prospective planning and processing.
A SNWs wheel of life starts with its creation: it could be defined, installed and executed non-automatically by humans, for example by using an inductive way, or by regular expressions. SNWs initiation [4] is one of the most important panoramas of this field, because it proposes high degree of algorithms execution and computerization. We can also enumerate on amalgam approaches that make end-user feasible to produce partially automatic SNWs by the mode of Graphical User Interfaces (GUI). For the drastically changing web content without any acknowledgment, SNWs perpetuation and support is one of the important areas for ensuring the continuous operation of SNWs systems.
SNWs are appropriate to the Social Media Networks extraction (SMNE) problem because unstructured HTML coded pages are presented in the format of syntactically structured. HTML is just a client-side presentation markup language, although SNWs can use HTML elements to conclude hidden information.

3. SOCIAL NETWORK WRAPPERS LIFE CYCLE
The life cycle of the in Figure 1 below consist of three phases' firstly creation, secondly initiation and lastly perpetuation and support.

3.1 Social Network Wrappers Creation
The following table 1 shows two categories of SNWs creation; Firstly is *Semi-Automatic Social Networks Wrappers (saSNWs) (those wrappers that could be described and implemented with human interaction)*, and Secondly *Automatic Social Networks Wrappers (aSNWs)* (those wrappers that requires no human interaction). Given below the table that classifies the various techniques required for the generation of the above mention types of wrappers.

The flowchart in figure 1 below depicts the process-cycle for the SNWs.

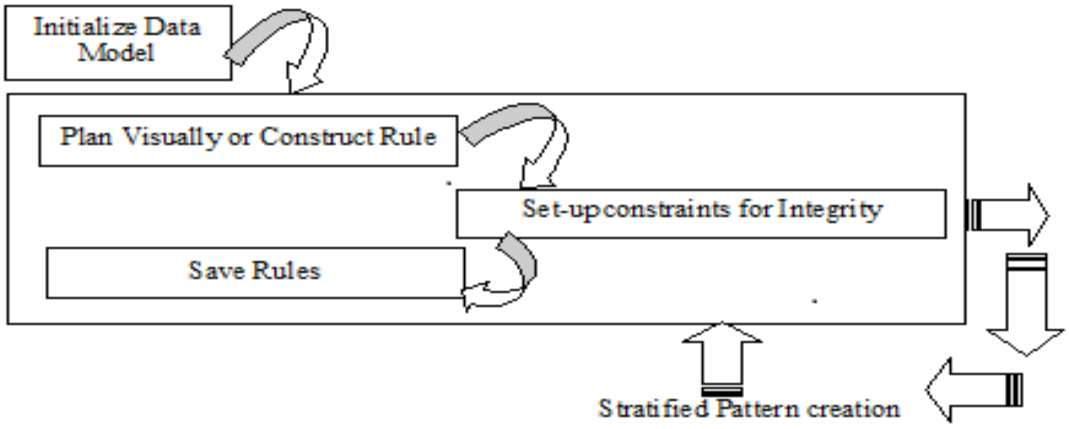

Figure 1: Diagram of SNWs creation

Categories	Techniques	Definition
SaSNWs	Approaches Based on: Logically mapping	Tools based on this approach used the *wrapper programming language*, taking into account the fact that the web page is not only consist of text strings but a represent a tree-like structure the DOM (Document Object Model). Many researchers like Baumgartner et al. [5], Gottlob and Koch [6] developed their own wrapping programming language.
	Approaches Based on: Regular-Expression	In this technique, manually complex rules are made or the formal language is used to recognize strings or patterns in the unstructured text. Generally regular expressions used on the web pages will rely upon the few traits like HTML elements, tags, tables structures etc. The commonly used tool for this approach is W4F [7], adopting an annotation approach: instead of putting users facing the HTML code, W4F eases the design of the wrapper through a wizard that helps the end-users to identify and annotate elements on the web page directly.
	Spacial Representation Optical Extraction	This techniques uses the OCR (optical Character recognition) algorithm. This approach rely on X-Y cut OCR algorithm used by browser to extract the completely different approach called Visual Box Model (8,9). Those users who don't have much understanding of wrapper programming language can use this technique because in this user can take the web pages of their own interest and use the GUI to build or generate the automatic wrappers.
aSNWs	Automatic Pattern Matching	The basic idea for this process is working with two different HTML web pages at the same time, in order to discover similar as well as dissimilar patterns between structures and content of pages. An good example of automatic wrapper generator is RoadRunner (10, 11).
	PTA(Algorithm for Partial Tree)	The basic foundation for this technique is that, the web page consist of data record regions (closed regions of the page). The partial tree algorithm will extract these regions by using matching pattern approach called as tree edit distance. Researchers Zhai and Liu (12, 13) had developed a Web data extraction system based on the partial tree algorithm technique.

Table 1: List of approaches for the saSNWs and aSNWs. [1]

3.2 Social Network Wrappers Initiation

In wrapper Initiation process, Extraction Rules (ER) are inferred from the training sessions and then applied to data extracted from Web pages. Many of the researchers used wrapper initiation techniques based on machine–learning approach which requires human involvement, domain expertise and needed huge amount of labeled Web pages given during training sessions.

WIEN [14] system has drawbacks of handling the missing values, it is based on the idea of coupling of excellent initiation learning techniques that enable the process to automatically label training web pages.

Hsu and Dung [15], developed SoftMealy, was the first initiation system designed for Web data extraction which uses bottom-up inductive learning approach to extract wrapping rules and depends upon non-deterministic finite state automata (State represent the data extracted and State transitions represent the rules for extraction) .

STALKER [16], given a supervised learning approach, in which, human intervention is required to place set of tokens on the web pages, identifying the information required for extraction with the capability of handling hierarchical structures, unordered items and the null values.

Statistical machine-learning-based systems were developed relying on conditional models [17] or adaptive search [18] as an alternative solution to human knowledge and interaction.

3.3 Social Network Wrappers Perpetuation and support

While SNW developing, irrespective of the approach applied to produce it, is only one problem which occurred during data extraction from social media networks (SMN): unlike static web HTML documents, Web pages drastically changes, evolve, and even their structure may also got changed, results to that SNWs cannot able to extract the data successfully. The most important phase of the online social media networks (SMNES) extraction system is the SNW perpetuation or support: this can be achieved humanly, modifying the SNW every time online pages alter; this techniques could do well for minor problems, but is not able to work successfully if the number of online pages is increased (for example, if an extraction process encounters thousands of pages, rapidly produced and frequently modified).

Kushmerick [19] described the SNWs authentication difficulty and, briefly, a few of human-coded SNWs maintenance approaches were defined to overcome simple problems. We had explored a viable practice presented in literature to automatically resolve the problem during SNWs support, called as logical-design-guided SNWs support.

Meng et al. [20] developed the Logically-Design-Guided SNWs support for online extraction of data starting from the observation that, alteration in online pages, even vast, always safeguard semantic traits (i.e. syntactic properties of dataset likes string lengths, data patterns etc.), annotations (descriptive information of the web page) and hyperlinks.

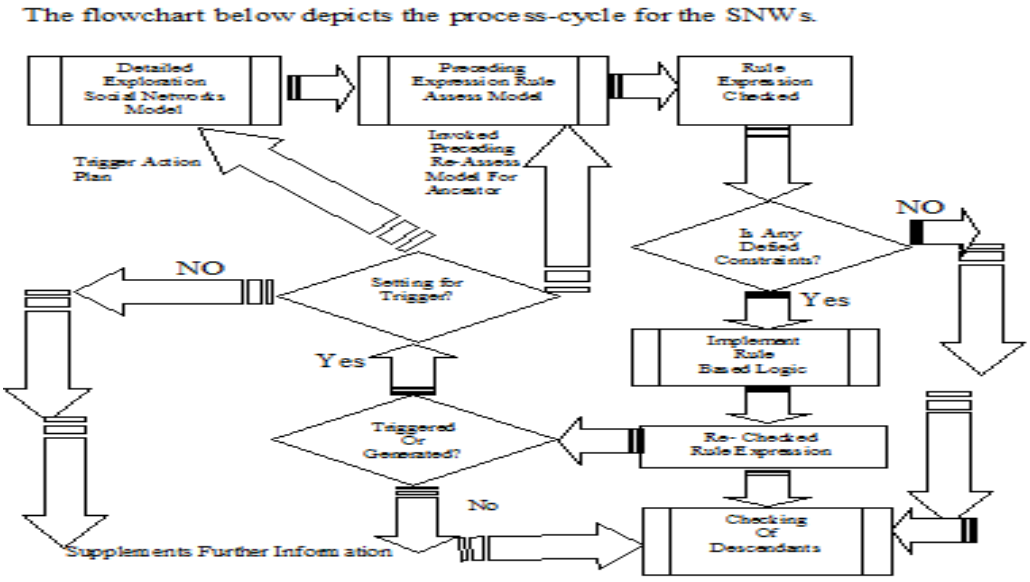

Figure 2: Diagram of SNWs Implementation, Perpetuation and Support

4. CONCLUSION AND FUTURE SCOPE

The approach discussed in the above paper is the Social Network Wrappers (SNWs), which are the basic foundation on which the data extraction process depends upon, from the online Social Media Networks (SMN). In our research paper, we discuss in detail about the SNWs, its process-cycle and the problem concerning with the SNWs maintenance. The details regarding our algorithmic and technical solution to this problem would be the future part of discussion. In future, we will discuss an innovative framework of SNWs adaptation without human interaction that will contributes to the state-of-the-art in this field.

5. REFERENCES

[1] Sarawagi, S.: Information extraction. Foundations and trends in databases 1(3), 261-377 (2008).

[2] Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys 34(1), 1{47 (2002).

[3] Kaiser, K., Miksch, S.: Information extraction. a survey. Tech. rep., E188 - Institute fur Softwaretechnik und Interaktive System; Technische University at Wien (2005).

[4] Kushmerick, N.: Wrapper induction for information extraction. Ph.D. thesis, University of Washington (1997).

[5] Baumgartner, R., Flesca, S., Gottlob, G.: Visual web information extraction with lixto. In: Proc. of the 27th International Conference on Very Large Data Bases, pp. 119-128.

[6] Morgan Kaufmann Publishers Inc. (2001). Gottlob, G., Koch, C.: Logic-based web information extraction. ACM SIGMOD Record 33(2), 87-94 (2004).

[7] Sahuguet, A., Azavant, and F.: Building light-weight wrappers for legacy web data-sources using w4f. In: Proc. of the 25th International Conference on Very Large Data Bases, pp.738-741. Morgan Kaufmann Publishers Inc. (1999).

[8] Gatterbauer, W., Bohunsky, P.: Table extraction using spatial reasoning on the css2 visual box model. In: AAAI '06 Proc. of the 21st national conference on Artificial intelligence, pp. 1313-1318. AAAI Press (2006)

[9] Krupl, B., Herzog, M., Gatterbauer, W.: Using visual cues for extraction of tabular data from arbitrary html documents. In: Special interest tracks and posters of the 14th international conference on World Wide Web, pp. 1000-1001. ACM (2005).

[10] Crescenzi, V., Mecca, G.: Automatic information extraction from large websites. Journal of the ACM 51(5), 731-779 (2004).

[11] Crescenzi, V., Mecca, G., Merialdo, P.: Roadrunner: Towards automatic data extraction from large web sites. In: Proc. of the 27th International Conference on Very Large Databases, pp. 109-118. Morgan Kaufmann Publishers Inc. (2001).

[12] Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: Proc. of the 14th international conference on World Wide Web, pp. 76{85. ACM (2005).

[13] Zhai, Y., Liu, B.: Structured data extraction from the web based on partial tree alignment. IEEE Transactions on Knowledge and Data Engineering 18(12), 1614-1628 (2006).

[14] Kushmerick, N.: Wrapper induction: efficiency and expressiveness. Artificial Intelligence 118(1-2), 15-68 (2000).

[15] Hsu, C.N., Dung, M.T.: Generating finite-state transducers for semi-structured data extraction from the web. Information systems 23(9), 521-538 (1998).

[16] Muslea, I., Minton, S., Knoblock, C.: A hierarchical approach to wrapper induction. In: Proc. of the 3rd annual conference on Autonomous Agents, pp. 190-197. ACM (1999).

[17] Phan, X., Horiguchi, S., Ho, T.: Automated data extraction from the web with Conditional models. International Journal of Business Intelligence and Data Mining 1(2), 194-209 (2005).

[18] Turmo, J., Ageno, A., Catal_a, N.: Adaptive information extraction. ACM Computing Surveys 38(2), 4 (2006).

[19] Kushmerick, N.: Finite-state approaches to web information extraction. Proc. of 3rd Summer Convention on Information Extraction pp. 77-91 (2002).

[20] Meng, X., Hu, D., Li, and C.: Schema-guided wrapper maintenance for web-data .In: Proc. of the 5th international workshop on Web information and data management,pp. 1-8. ACM (2003).