Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Social Media Networks (SMN) an Eye: To Envision and Extract Information

Dr. Mamta Madan
Professor Vivekananda Institute of Professional Studies, GGSIP University, India

Meenu Chopra
Assistant Professor Vivekananda Institute of Professional Studies, GGSIP University, India

Abstract: This paper tries to portrait outline study on the research problem of how to extract valuable knowledge from the various social media networks (SMN), in this regard many technologies, methods and procedures have been developed. Firstly, we intend to discuss and track the complete profile of the current data extraction online tools called as Social Media Networks Extraction System (SMMES), because they focus on the social spectrum applications where the data extraction tools can be applied. Secondly, we will cover the techniques which are related to the operation of the various tools which are used for the data generation from the SMN. A special focus is given to all the difficulties which are related to obtaining knowledge from online web sources, in particular from Social Media Networks. Thirdly, we categorize different fields of applications where web information extraction techniques can be applied, concentrating specially, on social and enterprise applications.

Keywords: Social Media Networks, SMNES, Network wrappers, online data Sources

1. Social Media Networks Extraction Systems (SMNES)

We can generically define a social media network system as a chronological series of procedures [1]. From the above definition, we interpret the two aspects of the problem (i.e. the extraction from SMN) firstly, Interlinked Web Pages (These systems also support the extraction of information from dynamically generated Web pages, usually built at run-time as a consequence of the user request, filling a template page with data from some database. The other kinds of pages are commonly called static Web pages, because of their static content). Secondly, Genesis of a Wrapper (A SMNE system must implement the support for wrapper genesis and wrapper execution.

Baumgartner et al. [6] has given a definition for the SMN extraction system, as "a software extracting, automatically and recursively, information which is changing rapidly from online web pages, and finally storing the extracted information into a database or any other back-end application". According to this definition we are able to define the current view of the problem for the information extraction, firstly, *Mechanization (or Automation) and Planning (or Scheduling)* Secondly, D*ata transmutation (or Data Transformation)* and thirdly, *Utilization of the Extracted Data*.

The following table shows techniques which possibly can be used to resolve the difficulty in obtaining the information from SMN.

Technique	Description
Interlinked or Association with different online web pages	The first phase of a generic Web data extraction system is the association or interaction [8]. The online social media sources, generally have static or dynamic pages, but also as RSS/Atom feeds [9], Microformats [10] etc., could be seen by the end-users, either in presentation or string mode. The present state can be presented by the systems that can help the extraction of information from web-pages attained by deep Web navigation [11].
Genesis of Wrapping	Generally, we characterize the thought of the wrapper as a process used for the obtaining amorphous or unorganized data from a target source and modified them into organized information [16, 17]. For a SMNES it is mandatory to have support for wrapper production and its implementation.
Mechanization and Planning	The mechanization and planning of the extraction process from the web pages is crucial part in data extraction systems [14]. The capacity to make the macros and not only to implement the many states of the same process, as well as to imitate the number of user's click, forms filling and choosing many menus, sub-menus and the buttons, the aid for AJAX technology [15] to manage the asynchronous modification of the page, etc. are few important mechanization features.

Information Transmutation	The data from heterogeneous sources could be wrapped, which means using different wrappers and also, probably, varied composition of data obtained. Data Transmutation is the process between of retrieving and delivering which includes phases like data cleaning [13] and conflict resolution [12], the end-users get the resultant product in the form of homogeneous data under a common and unique resulting structure.
Utilization of the Pull-out Data	When the fetching phase is over, then accessed data will be bundled in the specific format, finally, the resultant package of information is ready for use. Last phase, is to deliver the package (structured Data) to a end application like a native XML DBMS, a RDBMS, a data warehouse, a CMS, etc

Table 1: List of Techniques to extract the data from online social media networks. [1]

2. Related Research

Many researchers have done authenticate analysis on the data extraction problem from the Social Media Network (SMN). In 2002, Laender et al. [1] bestowed an important survey, giving extact ecology to categorize Social Media Network (SMN) Extraction Systems. They brought in a set of benchmark and a subjective study of different SMN's tools. Flesca et al. [3] and Kaiser and Miksch [2] explained variety of tools, techniques and approaches.

Chang et al. [5] proposed a three-dimensional classification of the SMN Extraction Systems, depended on the process problems, methods used and intensity of computerization. Fiumara [4] utilized these benchmarks to categorize four novel tools that are also discussed in this paper. Sarawagi bring out an enlightening study on Extraction of Information [7]: At the latest, the study from Baumgartner et al. [6] is the most recent brief survey on the state-of-the-art of the discipline.

3. Social Media Networks Tools

According to Laender et al. [1], the below given table represents the nomenclature to segregate the systems, on the criteria of *wrappers generation techniques*.

Table 2: A taxonomy for characterizing SMNE tools. [1]

Tool	Description
NLP-based Tools (Natural Language Processing)	These techniques were innate to extract information [18, 20, and 27]. They were applied to solve the problem in Information Extraction (IE). Difficulties like the generation of facts from newspaper, resumes, speech transcriptions in forums, email messages.etc.
Modeling-based Tools	Having confidence on the decided set of natives to analyze with the framework of the given page, these tools can able to search one or more items or the objects in the given page identical the native items or objects [22, 24]. Expertise knowledge is required in this domain, although it is a better approach for the fetching the data from SMN based on arrangements and rapidly changing online pages.
HTML-aware Tools	Some tools rely on the intrinsic formal structure of the HTML to extract data, using HTML tags to build the DOM tree (e.g. RoadRunner [25], Lixto [23] and W4F [26]). Boronat [28] analyzed and compared performances of common Web data extraction tools.
Wrapper Induction Tools:	These tools generate rule-based wrappers, automatically or semi-automatically: usually they rely on delimiter-based extraction criteria inferred from formatting features [19].
Languages for Wrapper Development:	Before the birth of some languages specifically studied for wrapper generation (e.g. Elog, the Lixto Web extraction language [22] extraction systems relied on standard scripting languages, like Perl, or general purpose programming languages, like Java to create the environment for the wrapper execution.

Laender et al. [1] also had given other *qualitative criteria* to classify the SMNES.

- ***Repercussion and Adjustment:*** Sources which are present online generally get modified without giving any premonition, for example the rate of modifications is not known before hand and therefore extraction systems that produce wrappers with a huge magnitude of repercussion exhibit good functioning. Also the Adjustment of the wrapper, transition from a one online source of data to another source, having equivalent domain, is another plus point.
- ***Usage Simplicity:*** Now days, Graphical User Interface (GUI) is a mandatory for present extraction tools. Technologies like Lixto[6], Denodo[1], Kapow Mashup Server, WebQL[2], Mozenda[3], Visual Web Ripper[4], to produce wrappers, merging with Web browsers, WYSIWYG editor interfaces etc. all are using rich user capabilities..

- ***Contents on Online Sources:*** Subjected matter or the text on the web can be broadly divided into two formats: Unstructured Data and Semi-Structured Format Data. The first one is adaptable with NLP-Dependent tools and Ontology-Dependent tools, the latter with the others.
- ***XML output:*** According to WWW[5], it is a linguistic industry standard for portraying web data. Nowadays, for commercial software, it is necessity to produce the fetched data in XML format.
- ***Intensity of Computerization:*** Gone those days when we use to fetch the data with human effort. Backing for complicated objects: Today, the online content is rich, dynamic and complex.
- ***Platform for Non-HTML sources:*** NLP-based tools apt for this domain, reason, a large quantity of data is saved online are in format of semi-structured text (emails, documentations, logs, etc.).

3. Classification Of Applications Of Social Media Extraction Techniques

Generating the data from heterogeneous sources from the web is the most crucial step of the data extraction process, reason; it is the essential to create a strong base of authenticate syntactic information. Whereas, Web 2.0 with features like rich client technologies, the consumer as producer philosophy etc. enhance the humans way to absorb the Web with Social Media Networks. Therefore, innovative progress put further pressure on the researchers to modify or recompile the rules to extract data from the Web sources and also explore the concept of applications related to Web.

The motive of this Section is to study and review few possible applications that are rigorously mutually dependent with Web data extraction processes. The following table depicts a taxonomy in which, we divided the key application fields into two categories one is Business or Organizational applications and second one Social or Communal applications, these applications massively do data extraction from Web sources.

Category	Examples
Business Applications	• Business Intelligence and Competitive Intelligence[13] • Customer care(NLP-based techniques are the best for them) • Context-aware Broadcasting[7] • Database creation(this is a key concept in the Web marketing sector) • Software Engineering[(the concept of rich internet applications(RIA[29]))
Communal Applications	• Social Media networks • Citation databases[10,11,12] • Comparison shopping[14,15] • Perception sharing[32] • Social bookmarks[33,34]
Prospective Applications	• Web harvesting[8,9] • Bio-informatics and Scientific Computing[30,31]

5. Conclusion and Future Scope

In this paper we concisely considered the present perspective that generally take care of the techniques and various online data mining sources, specially focus on the area analogue to SM (social media) networks and their data extraction. The aim of this research paper is to provide an overview about the Social Media Networks Extraction Systems and their field of applications particularly social applications. The process called network wrappers are the foundation on which the data extraction from Social Media Networks will depend upon and the detail study of our algorithm design, scientific and technical explanation related to the automated networks wrappers support are the future scope of discussion

WEB REFERENCES

1. www.denodo.com/
2. http://www.ql2.com/
3. http://www.mozenda.com
4. http://www.visualwebripper.com/
5. http://www.w3.org/
6. http://www.lixto.com/
7. http://www.appliedsemantics.com/
8. http://www.pubmed.com/
9. http://www.search-computing.it/
10. http://citeseer.ist.psu.edu/
11. http://www.informatik.uni-trier.de/ley/db
12. http://www.harzing.com/pop.html
13. http://www.sap.com/

14. http://www.qualityunit.com/unitminer/
15. http://www.bget.com/

REFERENCES

[1] Laender, A.H.F., Ribeiro-Neto, B.A., da Silva, A.S., Teixeira, J.S.: A brief survey of Web data extraction tools. ACM Sigmod Record 31(2), 84-93 (2002).

[2] Kaiser, K., Miksch, S.: Information extraction. a survey. Tech. rep., E188 - Institut f□ur Softwaretechnik und Interaktive Systeme; Technische University at Wien (2005)

[3] Flesca, S., Manco, G., Masciari, E., Rende, E., Tagarelli, A.: Web wrapper induction: a brief survey. AI Communications 17(2), 57-61 (2004)

[4] Fiumara, G.: Automated information extraction from web sources: a survey. Proc. of Between Ontologies and Folksonomies Workshop in 3rd International Conference on Communities and Technology pp. 1-9 (2007)

[5] Chang, C.H., Kayed, M., Girgis, M.R., Shaalan, K.F.: A survey of web information extraction systems. IEEE Transactions on Knowledge and Data Engineering 18(10), 1411-1428 (2006).

[6] Baumgartner, R., Gatterbauer, W., Gottlob, G.: Web data extraction system. Encyclopedia of Database Systems pp. 3465-3471 (2009).

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

[8] Wang, P., Hawk, W., Tenopir, C.: Users' interaction with world wide web resources: an exploratory study using holistic approach. Information processing & management 36(2), 229-251 (2000)

[9] Hammersley, B.: Developing feeds with rss and atom. O'Reilly (2005)

[10] Khare, R., C_ elik, T.: Microformats: a pragmatic path to the semantic web. In: Proc. Of the 15th international conference on World Wide Web, pp. 865-866. ACM (2006)

[11] Baumgartner, R., Ceresna, M., Ledermuller, G.: Deepweb navigation in web data extraction.In: Proc. of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, pp. 698-703. IEEE (2005).

[12] Monge, A.E.: Matching algorithm within a duplicate detection system. IEEE Data Engineering Bulletin 23(4) (2000).

[13] Rahm, E., Do, H.H.: Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin 23(4) (2000).

[14] 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).

[15] Garrett, J.J.: Ajax: A new approach to web applications. Tech. rep., Adaptive Path (2005). URL http://www.adaptivepath.com/ideas/essays/archives/000385.php.

[16] Irmak, U., Suel, T.: Interactive wrapper generation with minimal user e_ort. In: Proc. of the 15th international conference on World Wide Web, pp. 553-563. ACM (2006)

[17] Zhao, H.: Automatic wrapper generation for the extraction of search result records from search engines. Ph.D. thesis, State University of New York at Binghamton (2007). Adviser-Meng, Weiyi.

[18] Berger, A., Pietra, V., Pietra, S.: A maximum entropy approach to natural language processing. Computational linguistics 22(1), 39-71 (1996).

[19] Anton, T.: XPath-Wrapper Induction by generalizing tree traversal patterns. MIT Press (2004).

[20] Manning, C.D., Schcutze, H.: Foundations of statistical natural language processing. MIT Press (1999).

[21] 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 Data Bases, pp. 109-118. Morgan Kaufmann Publishers Inc. (2001).

[22] Gatterbauer, W., Bohunsky, P., Herzog, M., Kr□upl, B., Pollak, B.: Towards domain independent information extraction from web tables. In: Proc. of the 16th international conference on World Wide Web, pp. 71-80. ACM (2007).

[23] 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.Morgan Kaufmann Publishers Inc. (2001).

[24] Embley, D., Campbell, D., Jiang, Y., Liddle, S., Lonsdale, D., Ng, Y., Smith, R.: Conceptual-model-based data extraction from multiple-record web pages. Data & Knowledge Engineering 31(3), 227-251 (1999)

[25] Baumgartner, R., Ceresna, M., Ledermuller, G.: Deepweb navigation in web data extraction. In: Proc. of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, pp. 698{703. IEEE (2005).

[26] Sahuguet, A., Azavant, 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).

[27] Winograd, T.: Understanding natural language. Cognitive Psychology 3(1), 1-191 (1972).

[28] Boronat, X.: A comparison of html-aware tools for web data extraction. Master's thesis, University at Leipzig, Fakultat for Mathematic und Informatic (2008). Abteilung Datenbanken.

[29] Amal_tano, D., Fasolino, A.R., Tramontana, P.: Reverse engineering _nite state machines from rich internet applications. In: Proc. of the 15th Working Conference on Reverse Engineering, pp. 69-73. IEEE (2008).

[30] Gatterbauer, W.: Web harvesting. Encyclopedia of Database Systems pp. 3472-3473 (2009)

[31] Weikum, G.: Harvesting, searching, and ranking knowledge on the web: invited talk. In: Proc. of the 2nd International Conference on Web Search and Data Mining, pp. 3-4. ACM (2009)

[32] Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classi_cation of product reviews. In: Proc. of the 12th international conferenceon World Wide Web, pp. 519{528. ACM (2003)

[33] Quattrone, G., Capra, L., De Meo, P., Ferrara, E., Ursino, D.: E_ective retrieval of resources in folksonomies using a new tag similarity measure. In: Proc. of the 20th Conference on Information and Knowledge Management, pp. 545-550. ACM (2011)

[34] Quattrone, G., Ferrara, E., De Meo, P., Capra, and L.: Measuring similarity in large-scale folksonomies. In: Proc. of the 23rd International Conference on Software Engineering and Knowledge Engineering, pp. 385{391 (2011)

