NEW EXTRACTION SYSTEM
STRUCTURED INFORMATION ACQUIRED FROM MULTI-SOURCES

Naser F. M. EL-firjani, Ebitisam K. Elberkawi
Faculty of Information Technology, Benghazi University, Libya
nasser195it@yahoo.com

الملخص

كثير هي المستندات ومواقع الويب والوثائق الطبية والتعليمية والقانونية وخلافه من كل المجالات التي تحوي عديد من المعلومات في صورة تجعل من صانع القرار في مهمة صعبة لعدم خروج هذه المعلومات بالصورة المفيدة والمنظمة التي تسهم في اتخاذ قرار صحيح . ومن هنا جاءت أهمية توفر تقنية تكون فعالة لأداء هذه المهمة وذلك بتحليل المعلومات أيا كانت صورتها أو ترتيبها لكي يستخلص منها معلومات مرتبة وفعالة في عملية اتخاذ القرار.

تهدف هذه الدراسة الى تصميم نظام استخراج المعلومات (IES) حيث يمكن لهذا النظام أن يقوم بتجميع معلومات مختلفة حول امر محدد من عدة مواقع ويب ومن ثم تحليل المعلومات المجمعة لإنتاج معلومات في صورة مفيدة.

واختير كا حالة دراسة نظام تجميع يساعد المستخدم في اتخاذ القرار لاختيار السيارة المرغوب اقتناءها بماركة محددة من عدة مواقع ويب وذلك بالبحث في عدة مواقع حول الماركة المطلوبة واستخراج المعلومات الهامة والمفيدة من كل تلك المواقع. وبنهاية الدراسة تم طرح التوصيات للدراسات المستقبلية حول هذا النظام.

Abstract

Textual information such as news, government documents and medical alerts are available in many repositories and transmitted in an unstructured form which has made information retrieval become a difficult task to achieve. Hence, there required an effective and efficient technique for the extraction and analysis of this information in a structured form. This study designed

information Extraction System (IES) that can extract unstructured information from many websites and analyze the information in a structured form. The architecture, benefits and recommendation for future studies on this system are provided as well.

Keywords: Information Systems; Information Extraction; Extraction Methods.

I. INTRODUCTION

Recently, the availability of textual information such as the news, government documents, medical alerts and records in digital form is on the increase in information repositories such as the internets and the intranets [1]. These information are transmitted through an unstructured, free text documents contain weak-related variant information thus becomes a difficult task to search in the repository and difficult to interpret and reason, thereby required the need for an effective and efficient technique for the analysis of a free-text and valuable information discovery in the form of a structured information which has resulted in the development of IES [1], [2] &[3]. IES aims at identifying a set of predefined information in a certain domain by neglecting the irrelevant information within the domain in any of the domains, including text mining, automated website annotation, business intelligence and information management [1], [4] &[2].

The concern of information extraction is to derive factual structured information from the unstructured text [1]. For example, consider ones interest in extracting identifying information about the main actor, location and the number of the victims of violence from online news. In extracting this information, the identification of a certain small scale of structure like noun phrases that denoted the person or the group of people involved in the event, geographical location, the numeral expressions and the semantic relations between them are required to enhance the extraction process [1]. Thus, information extraction task is to identify a particular class of entities in the event, the relationships and the events in the natural language texts and the extraction of the relevant class of objects or relationships in the events [1], [4], [5]

&[6]. Therefore, this study aims at presenting the concept of and developing an IES by explaining the architecture and the benefits of the system developed.

The paper organized as follows: Section 2 contains task and sub-tasks of the IES. Benefits of the IES considered in this paper are presented in Section 3. Section 4 discusses in detail the architecture of the proposed system. Finally, conclusion and future recommendation are presented in Section 5.

II. TASK AND SUB-TASKS OF THE IES

As can be seen in figure 1, the task of Information Extraction (IE) is to identify a predefined set of concepts in a specific domain, ignoring other irrelevant information,
where a domain consists of a corpus of texts together with a clearly specified information need. In other words, IE is about deriving structured factual information from unstructured text. The extracted information is pre-empted in a template (or objects), which consists of a number of attributes (slots), which are to be instantiated by an IE system as it processes the text.

The pre-specified attributes are often strings from the events or texts, one among the pre-specified values or a reference to an object template that has been created previously. Information Extracted system can be thought of as database that contains object population, since it enhances the creation of database structured representation of selected certain information taken from the analyzed text [1].

The general aim of IES is to create an easily machine-readable text for sentence processing. The following are the subtasks undertaken by an IES:

Named Entity Extraction which include: known entity name recognition (for individual people or institution), places and certain types of numerical expression by employing the existing knowledge of the domain for the extraction exercises.

Co-reference Resolution: to detect a conference and anaphoric links between texts and entities.

Extraction of Relationship: to identify the relationship between entities, such as a person and the organization where he/she works and also between a person and its location.
Extraction of Semi-Structured Information: such as tables and comments.
Language and Vocabulary Analysis: this consists of terminology extraction and analysis, which relates to finding and giving corpus to relevant terms
Audio Extraction: finding, extraction and analyzing audio based templates for a particular repertore.

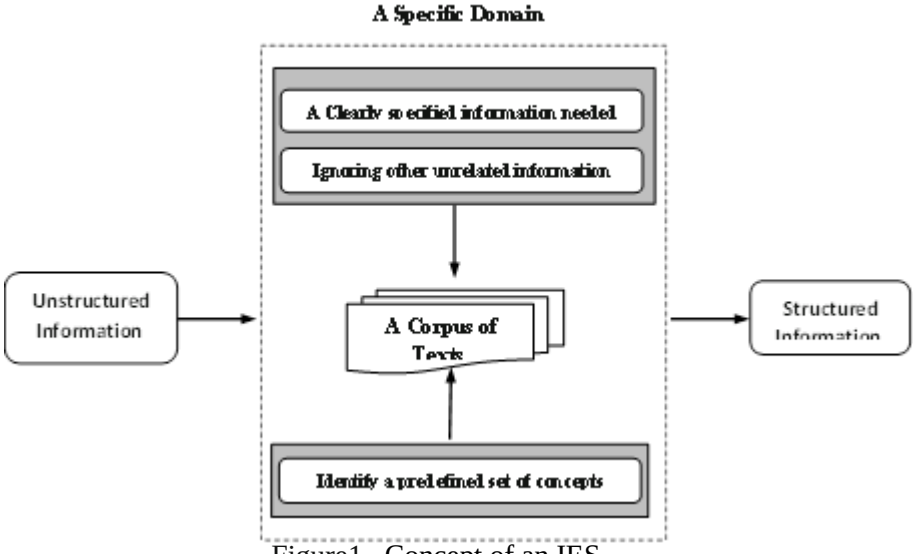

Figure1. Concept of an IES

III. BENEFITS OF THE IES

IES is greatly valued by the commercial world of business largely due to its easy adoption, understanding, debugging, interpretation and maintenance in the face of change in requirements [9] & [10].

Information extraction has also been very significant in cutting down the labour cost for developing or adapting extractors and the cost of hardware of computer resources needed for a particular business problems. This has been valued as researchers have well

developed sense of labor cost of writing extraction rules are tedious and not practicable [9].

IV. ARCHITECTURE OF THE PROPOSED SYSTEM

Although IE systems built for different tasks may differ from each other significantly, there are certain core components shared by most IE systems and the Flexible IE systems that turn Web pages into program-friendly structures [11].Users seeking information on cars online need to browse a lot of sites to get latest information of the car brand. Thus, this study developed a simple IES that helps users to extract information about car brand from many websites. In this case, the source of the information is a web server. The wrapper which is a system component that provides a single uniform query interface that access many sources of information. The wrapper queries the web server to gather information about the query using the HTTP protocols. Then, implement the information extraction process on the content of the HTML documents, followed by the integration of the information extracted with the other sources [7]. The extraction system is explained task and sub-tasks of IES above.

The IES is designed as follows:

Content management framework (CMC) which is written with PHP language, CSS and use object-oriented programming techniques were used to design system.Rich site summary (RSS) was used as

an approach to depict information about cars on any web contents that can be available for "feeding" (syndication or distribution) from a publisher that is online to the users of the web.

RSS is an Extensible Markup Language (XML) application that abide to the Consortium of World Wide Web (RDF) Resource Description Framework. Primarily developed by Netscape for the channels of Netcenter browser. The specification of RSS now can be available for everyone to use [8].

RSS feeding was used as the approach of extracting information from the designed system.

The technicalities employed in this study is copying the RSS link from the websites that information is extracted to the field

designated for showing information in the designed system (website).

The goal of this system is to extract the latest information on car brand from different available websites. The system shows the extracted result of car brands in different fields. This website will help the user to specifically search as many as possible information and collate the information about the searched car brand. The interface of the system is very simple to use. This simplicity will help the users to use the system easily.

Some screenshots about the system are presented in figures 2,3:

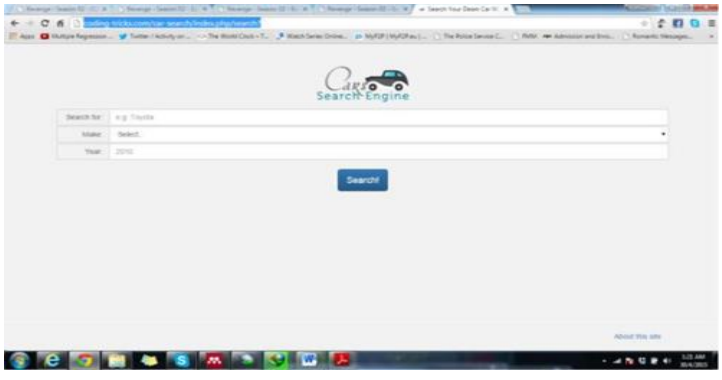

Figure 2. Interface of the system for information extraction on cars

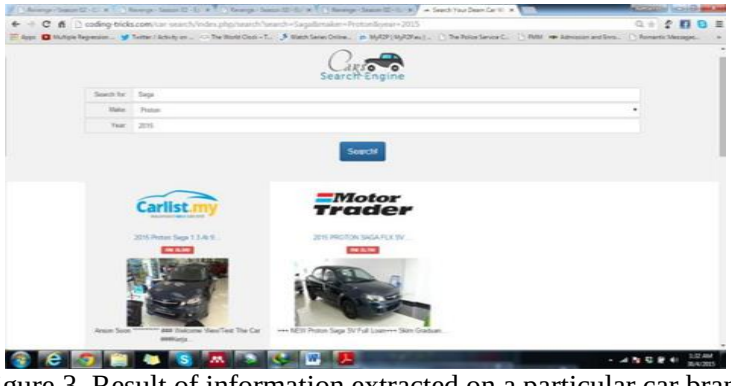

Figure 3. Result of information extracted on a particular car brand (Proton, Saga, 2010)

V. CONCLUSION AND FUTURE RECOMMENDATIONS

This system is expected to enable users in obtaining structured information of car brands from different websites available online. The system is developed to meet the needs of large numbers of users as possible. The emergence of new technologies in hardware and software make programmers and developers to work professionally in designing a system that is easy to use. However, the development of most IESs is based on training by human interpreted corpora, however, the construction of the corpora for information extraction accuracy is a burdensome task [6].

One better way is the use of an active learning to enhance the reduction of the required number of learning by the system and this required the selection of the most informative texts to give human annotators. Therefore, more research is required in this aspect to reduce the supervisory demand of the training of the IES. Also, there is a need to develop an unsupervised leaning method to eliminate the supervision requirement of the IES.

REFERENCES

[1] T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, “Multi-source, Multilingual Information Extraction and Summarization,” Theory Appl. Nat. Lang. Process., pp. 23–50, 2013.

[2] S Singh, "Natural Language Processing for Information Extraction" arXiv preprint arXiv:1807.02383, 2018 - arxiv.org

[3] M. Abdelmagid, A. Ahmed and M. Himmat "Information Extraction methods and extraction techniques in the chemical document's contents: Survey,", ARPN Journal of Engineering and Applied Sciences, vol. 10, pp. 1068-1073, 2015.

[4] M. Kayed, M. R. Girgis, and K. F. Shaalan, “A Survey of Web Information Extraction Systems,” IEEE Trans. Knowl. Data Eng., vol. 18, pp. 1411–1428, 2006.

[5] M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, “Open language learning for information extraction,” EMNLP-CoNLL ’12 Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn., pp. 523–534, 2012.

[6] R. J. Mooney and R. Bunescu, "Mining knowledge from text using information extraction," ACM SIGKDD Explor. Newsl., vol. 7, no.1, pp. 3–10, 2005.

[7] J. Tang, M. Hong, D. Zhang, B. Liang, and J. Li, "Information Extraction: Methodologies and applications," Emerg. Technol. Text Min. Tech. Appl., pp. 1–33, 2008.

[8] J. G. Hendron, "RSS for educators: blogs, newsfeeds, podcasts, and wikis in the classroom," pp. 1–3, 2008.

[9] L. Chiticariu and F. R. Reiss, "Rule-based Information Extraction is Dead ! Long Live Rule-based Information Extraction Systems !," Emnlp, no. October, pp. 827–832, 2013.

[10] G. Fette, P. Kluegl, M. Ertl, S. Störk, and F. Puppe, "Information
Extraction from Echocardiography Records," Work. Notes LWA 2011 - Learn. Knowledge, Adapt., 2011.

[11] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled
Shaalan, "A Survey of Web Information Extraction Systems" IEEE Transactions on Knowledge and Data Engineering · October 2006.