Standardization of Unstructured Textual Data into Semantic Web Format

A Mini project Report
submitted in partial fulfilment of
the requirements for the award of the degree of

Bachelor of Technology
in
Computer Science and Engineering
(University of Calicut)
by

Fasna P.P.(MCS09303310)

Department of Computer Science & Engineering
MES College of Engineering, Kuttippuram
(ISO 9001 : 2000 Certified Institution, Affiliated to University of Calicut)
Thrikkanapuram PO, Malappuram Dt, Kerala - 679573
2011-12

Certificate

*This is to certify that the mini project report entitled "***Standardization of Unstructured Textual Data into Semantic Web Format***" is a bonafide record of the work done by* **Fasna P.P.** *(Reg no: MCS09303310), under our supervision and guidance. The report has been submitted to the* **Department of Computer Science and Engineering** *of* **MES College of Engineering, Kuttippuram** *in partial fulfilment of the award of the Degree of* **Bachelor of Technology in Computer Science and Engineering, during the year 2015-16.**

Dr. P.P. Abdul Haleem
Professor and Head
Dept.of Computer Science and Engineering
MES College of Engineering

Mr K.A. Abdul Nazeer
Project Guide
Assistant Professor
Dept.of Computer Science and Engineering
MES College of Engineering

Abstract

Analysis done on the nature of the data posted on the Net reveal that more than 80% of the data over the Net is in unstructured text format. Hence extracting information from text is of paramount importance both for academic and business purposes. Simultaneously, evolution of web technology led to the novel concept of Semantic Web, which is an extension of the current web in which information is given well-defined meaning, enabling computers and people to work in cooperation in a better way. Integration of voluminous, legacy text data that are unstructured and semistructured, into Semantic Web format is a challenging and daunting task for the research community. This thesis work is an attempt to marry the concept of Semantic Web format with unstructured text, thus to enable the computers, the discovery of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information semantically together so that it can be directly used by the Semantic Web projects. The design of the proposed system is found to be complex, due to the primary fact that the work encompasses several research areas and these results are still in evolving stage.

Acknowledgements

Your acknowledgements

Fasna P. P.

Afeef Ali P. P.

Faheem P. P.

Saja P. P.

Contents

Figures

Figure

Chapter 1

Introduction

To date, the Web has been developed most rapidly as a universal medium of documents for people rather than for data and information that can be processed automatically. The essential property of the WWW is its universality. The power of a hypertext link is that "anything can link to anything".

Amazing growth of the internet resulted in large scale accumulation of information, making it difficult for humans to understand. It is estimated that more than 80% of data existing over the Internet and Intranet within organizations are in the form of unstructured text. Hence a considerable amount of human hours is spent in ineffective searches through multiple information sources including web sites and other conventional sources. This problem of information overload is further worsened due to the unstructured format of the content.

The challenge of the Semantic Web, therefore, is to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web. The Semantic Web is not a separate Web but an extension of the current one [**?**], in which information is given well-defined meaning, enabling computers and people to work in cooperation, in a better way. The first steps in weaving the Semantic Web into the structure of the existing Web are already initiated. In the near future, these developments will usher in significant new functionalities as machines become capable to process and "understand" the data, than they merely

display at present.

1.1 Adding Figures

1.2 Adding Tables

Table 1.1: Features of XML and YAML

Feature	**XML**	**YAML**
Read/Writeability	Human Readable	Human Readable
Data Exchange Among Applications	Applicable	Applicable
Structured Data	Yes	Yes
Intrinsic Support for Data types	No	Yes
Verbosity	High	Less
Number of APIs	High	Less
Platform Independence	Yes	Yes
Schema Aware	Yes	Yes
Built-in Schema	Yes	No
Message Level Security Enhancements	Yes	No
Schemes for Prevention of Rewriting Attacks	Yes	No

1.3 Adding Equations

Here is a displayed

$$\int \frac{d\theta}{1+\theta^2} = \tan^{-1}\theta + C$$

equation.

An equivalent way to format the same displayed equation is to spell out the delimiters in words.

Here is a displayed

$$\int \frac{d\theta}{1+\theta^2} = \tan^{-1}\theta + C$$

equation.

A variant of the above is the following.

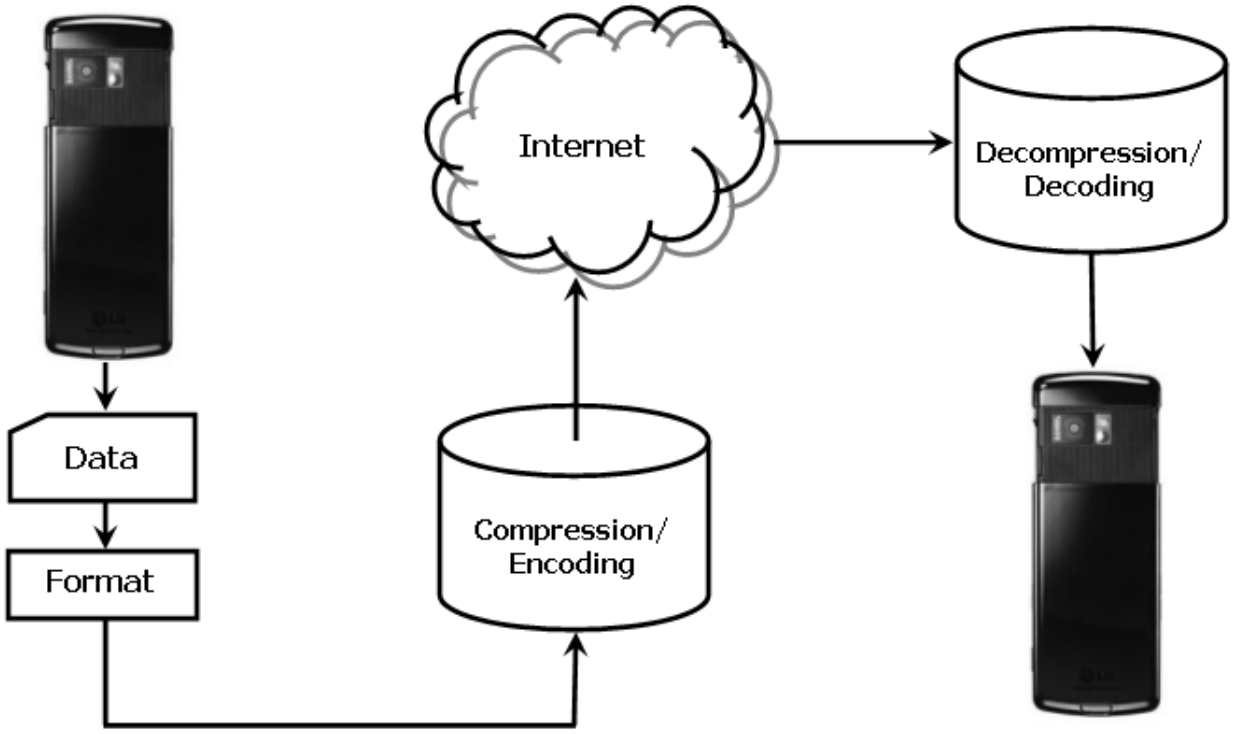

Figure 1.1: Verbose Data Formats using Compression/ Decompression

Here is a displayed

$$\int \frac{d\theta}{1+\theta^2} = \tan^{-1}\theta + C \tag{1.1}$$

equation.

There is a difference between the "displaymath" environment and the "equation" environment: the latter automatically typesets a formula number.

What if you want to refer to a numbered equation by number? How do you manage this if the equation number is automatically generated? LaTeX has a simple nd **??**echanism for handling symbolic cross references. Here is an example.

The formula

$$E = mc^2 \tag{1.2}$$

has passed into popular culture, but the true significance of the mass-energy equation (1.2) is ...

This paper centers around the humble intention to find out whether it is possible to unearth the hitherto unknown information from the unstructured text jungle over the WWW, and present it as a structure in conformity to the model of Semantic Web.

This reoport is organized as follows: Chapter **??** outlines the literature survey. Chapter **??** is about the problem definition. In Chapter **??**, an overview of the proposed system is explained. Chapter 2 contains conclusion and a breif discussion about the future work.

Chapter 2

Conclusion and Future Work

This paper described the standardization of unstructured text into Semantic Web Format, which is achieved in four stages.

Considerable human hours are wasted in the repetitive task of interpretation and semantic annotation to reclaim the knowledge implicitly conveyed in the vast amount of ever growing available text content. This is due to the fact that the major part of the implicit semantic knowledge is not taken into account by state-of-the-art information access technologies like search engines, which restrict their indexing activities to superficial levels, mostly the keyword level. The work reported in this thesis becomes significant under these circumstances.

The process of unearthing information from text is extremely complex, considering the unpredictability of information packed in unstructured text. Hence a semi automatic method based on self learning mechanism is presented.

An important feature of such a system, is to have a facility that gives user the control over the process while automatically taking care of the extraction process based on the vocabulary base. This system is equipped with this feature in the sense that at the extraction phase it always performs the full extraction, based on the settings defined by the user.

Judging the correctness of the pre-selection recommended by the user, is a hard task in the process of semantic information extraction. To minimise the difficulty of this task, text highlighting of the annotated information was done in

the original document.

To make things more deterministic for the user, for the next generation system, it is suggested to have a text highlighting of the extracted information in the original document. The user could, by clicking on each of the suggestions, see the extracted highlighted entity within a context and more easily determine its correctness. Also, extraction techniques have to be more effective to minimise user intervention, thus effectively supplementing the system automation.

Although present work is just confined to the text, no one can ignore the luring and challenging task of effective acquisition, organization, processing, sharing, and use of the knowledge embedded in multimedia content as well as in information and knowledge based work processes.

2.1 Example

yuyuef byefb dfygfduy n]

dfgdsfnjkk

Bibliography

[1] "Kwalify," Available at http://www.kuwata-lab.com/kwalify/.

[2] O. Ben-Kiki, C. Evans, and I. dt Net, "YAML aint́ markup language (YAML) version 1.2, 3rd edition," January 2010, Available at http://www.yaml.org/spec/1.2/spec.html.

[3] Y. Han, X. Yang, P. Wei, Y. Wang, and Y. Hu, "ECGSC: Elliptic curve based generalized signcryption," Lecture Notes in Computer Science, vol. 4159, pp. 956 – 965, 2006.

[4] T. Bray, J. Paoli, Sperberg-McQueen, E. Maler, and F. Yergeau, "Attribute-value normalization," November 2008, W3C Recommendation. Available at http://www.w3.org/TR/REC-xml/AVNormalize.

[5] G. White, J. Kangasharju, D. Brutzman, and S. Williams, "Efficient XML interchange measurements note," July 2007, Available at http://www.w3.org/TR/exi-measurements/.

[6] S. Lanka and P. Parikh, "XML shredding," Midterm Project, New York University, Courant, November 2000.

[7] W. Chou, "Elliptic curve cryptography and its applications to mobile devices," Tech. Rep., University of Maryland, College Park, Maryland, 2003.

[8] N. Potlapally, S. Ravi, A. Raghunathan, and N. Jha, "Analyzing the energy consumption of security protocols," in ISLPED '03: Proceedings of the 2003 International Symposium on Low Power Electronics and Design, 2003, pp. 30 – 35.

[9] Y. Zheng, "Digital signcryption or how to achieve cost (signature & encryption) << cost(signature) + cost(encryption)," Lecture Notes in Computer Science, vol. 1294, pp. 165–179, 1997, Available at citeseer.ist.psu.edu/zheng97digital.html.

[10] Johnson and Menezes, "The elliptic curve digital signature algorithm (ECDSA)," Technical report CORR 99-34, Department of C&O, University of Waterloo, 1999, Tech. Rep., August 1999.

[11] M. Göksedef and Şule Gündüz-Öğüdücü, "Combination of web page recommender systems," Expert Systems with Applications, vol. 37, no. 4, pp. 2911–2922, 2010.

[12] A. Beach, M. Gartrell, X. Xing, R. Han, Q. Lv, S. Mishra, and K. Seada, "Fusing mobile, sensor, and social data to fully enable context-aware computing," in HotMobile '10: Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications. New York, NY, USA: ACM, 2010, pp. 60–65.

[13] R. Schmelzer and T. VanDersypen, XML and Web Services Unleashed. Indianapolis, IN, USA: Sams, 2002.

[14] H. R. Elliotte, XML Bible. New York, NY, USA: John Wiley & Sons, Inc., 2003.