Semantical Enhancement of Documents

Andrei-George Giurgiu

January 16, 2014

Abstract

In recent years, one of the most prominent areas of research has been the semantic enhancement of data. Some of the problems that the proponents of this field are trying to tackle are the vastness, the inconsistency and the uncertainty of the data. The ever-growing abundance of data poses a challenge to society and to all of the sciences (see [1]). The unusability of simple data by machines - a computer cannot understand data and make use of it as knowledge - is the reason for missing machine support with these issues. Computers are not equipped with the necessary tools to extract meaning from data, since it is not being currently provided in a form in which it can be easily understood by humans or at all by computers. A building block in successfully alleviating problems with machine assistance is the semantic augmentation of computer and human readable documents.

Text documents are rich data structures that, similarly to other families of documents, like presentations and spreadsheets, are used for delivering information and knowledge. Their full potential is, however, surprisingly underused. The content of a text document, in its simplest form, can be merely understood as data. However, communicating data is not enough - not for people nor for machines. In order for data to be useful, it needs to be understood by users. While humans are often able to convert the content of said documents to information and to map it to real world knowledge, computers are still lagging far behind. Without the existence of knowledge bases to query for meaning, data has no external meaning for a computer and content cannot be mapped to form. The main purpose of semantically enhancing text documents is enabling humans and computers to provide more than simple data and making it possible for them to understand content as information. What we aim to achieve through this research is to provide a framework that will be able to render text documents in a manner that enables fruitful exchange of information between humans and machines. A more knowledgeable machine is a smarter machine and can thus better assist human users. The way we aim to enable this enhanced transaction of information is by inserting meta-data for semantic objects like words, phrases, paragraphs, images, etc in text documents. The framework under development will be platform and application independent, while still relying on the open API of the invaded software.

References

[1] Andrea Kohlhase, *Semantic Interaction Design: Composing Knowledge with CPoint, Fachbereich 3 (Mathematik und Informatik) Universitat Bremen.* 2008.