Inspiratron.org - Natural language processing, machine learning and cybersecurity

What is the structure in data?

by Nikola Miloševi? - Saturday, June 21, 2014

https://inspiratron.org/blog/2014/06/21/structure-data/

Have you ever wondered what is structured data?

Why is this important? Well, from structured data we can extract semantic and know what is in the data, use the data. However, if we don't know how data is structured, we will be unable to extract semantics and to understand it. There are plenty of data available around us and on internet. However, most of this data is not very structured. Some data have some structures and can be mined, however, some cannot be so easily. I would like to analyze how data is structured in most commonly used way to present data - article. However, here I would not go trough any code or how things should be done. It will be more philosophical consideration what is structure for human and what is structure for machines and how it can be extracted. And mainly I would like to focus on what are for human structured data, like tables and lists.

x	$\sin x$	$\cos x$	$\tan x$	$\csc x$	$\sec x$	$\cot x$
$30° \equiv \frac{\pi}{6}$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{3}}{3}$	2	$\frac{2\sqrt{3}}{3}$	$\sqrt{3}$
$45° \equiv \frac{\pi}{4}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	1	$\sqrt{2}$	$\sqrt{2}$	1
$60° \equiv \frac{\pi}{3}$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	$\sqrt{3}$	$\frac{2\sqrt{3}}{3}$	2	$\frac{\sqrt{3}}{3}$

Structure of the article

When we take a look at some article, what structure can be seen. There are some. First of all, article has title. Title is separated from the rest of the text with some particular typeface and font size. Article can have couple of subtitles, which also has some other typeface and font size. Rest of the article has some text. However this text is also structured in paragraphs, each paragraph telling some unit of the story. On the first sight it seems like there cannot be taken more structure than that. However, it is wrong conclusion. Unstructured text has a lot of structure in it. Actually it is property of language that makes text and spoken language structured. Each language has a set of rules, mainly described in grammar, that help extract some structure from text. Using natural language processing tools such as part of speech taggers and dependancy taggers this structure can be seen. However, the question would be **WHAT IS STRUCTURE IN DATA?** The simplest answer would be that data is structured if anyone (even machine) can determine from data without any ambiguity what are all the relationships between entities and what are all the attributes of the entity. In text it is quite hard, but a lot of research is done over it and more and more structured data can be extracted from free text. However, articles may contain so called structured parts. Structured part of the text could be tables or lists. But are these structured items really structured?

Are structured things structured?

	Column Heading	
	Column Subheading	**Column Subheading**
Row Heading	Column Entry	Column Entry
Row Subheading	Column Entry	Column Entry
Row Heading	Column Entry	Column Entry

So here we come to the topic of my current research. It is obviously called structured data. However, they are only structured visually. In data and text mining we are looking for semantic structures. Same semantic structure or meaning of the data can be represented in the number of visual ways. One person may put some labels in the header of the table, while the other may put same data in the stub column. Third person may granulate data with sub-headers. Forth may make just a key value list. And there may be dozens of visual structured representation of the same data in table-like structures. This face makes visually structured data at least as hard to mine as it is unstructured text. Why at least as hard as free text? Free text is quite standardized with grammar. Grammar describes what can and what cannot be the

next word, how they semantically relate and so on. This does not exist with tables. Tables are not standardized at all. It is up to author to create table how he/she wants. Even thou, it is to some degree easier to determine what is related with what it is still hard to extract semantics. Previous sentence is not 100% correct, since with large number of headings, sub-heading, group headings it can be easy to lost even human reader. When we talk about human reader, most tables are very hard to read even for human readers. Patricia Wright in 1970' did a research where she showed how readers are more likely to make a mistake with larger number of dimensions of the table (or table's complexity). She also argues that people has to be trained in order to be able to read table correctly. In most cases people don't get the training to read the table, but the table is from their field or the field they can understand. So their background knowledge about the field helps them to identify relationships between entities in the table. However, this is hard to archive with machines even with domain resources and loads of available Linked Data.

Less structure in structured data

It can be concluded that visually structured data are actually less structured than unstructured text, because of lack of rules and because of needed background knowledge to resolve relationships. Both of these lacking things can be found in unstructured text. Relationship is most often explicitly said, while structure can be found in grammar. There is still a long way to go, but we hope we can resolve these issues in visually structured data.
