Modelling the Dynamics of Web Content

Matúš TOMLEIN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
matus@tomlein.org

The web content has a very dynamic nature, it frequently changes and spreads accross various information channels on the web. The behaviour of web content can be observed and analysed, however it requires a lot of archived data to be able to draw any conclusions. It can also be a challenging task to efficiently analyse the large amounts of data and recognise the ways in which the content changes and spreads accross websites.

On the other hand, the knowledge of the behaviour of web content is useful in many areas of software engineering. It can improve and optimize search algorithms for web content. Predicting when a website will change can be useful in web proxies that provide a cache of web objects. It can provide some basis for recommending similar content and also for prefetching websites.

The web content consists of different kinds of information that need to be recognised in order to process them. Some are less important and provide a lesser value to the user yet some are interesting and provide a real value. Both of these kinds of information can be present on a single website at the same time.

The user gets most value out of information that create the corpus or the main part of a website. That could be an article, a list of links to websites or any kind of relevant information.

Other computer generated content, like the number of visitors to a website, the current date and time or advertisements provide a less significant value to the user. In most cases it is a good idea to filter such content when analysing a website.

Distinguishing these kinds of information is also important when tracking changes of a website. Computer generated content and advertisements tend to change frequently, however these changes are usually not interesting to the user or to the analysis of the website.

Apart from changes in the main content of a website, there is another kind of content that is potentially interesting to observe, and that is visitor-generated content such as discussions or polls. This content can add additional information to the website that might be worth analysing.

---

* Supervisor: Jozef Tvarožek, Institute of Informatics and Software Engineering

Spring 2013 PeWe Workshop, April 5, 2013, pp. 91–92.

In terms of analysis of changes of web content, research has been done to recognise differences between two subsequent versions of an HTML document [2]. HTML represents elements as nodes in a hierarchy and although traditional algorithms for differencing text documents might work in HTML as well, it is advisable to employ an algorithm that can recognise changes in the hierarchy using tree comparison. Such algorithms can recognise insert, delete and move operations by differencing two HTML documents as can be seen on figure 1. To make the analysis aware of how content spreads accross documents, further improvements need to be made. This is an open challenge to come up with a method that can effectivelly detect the attributes of a dynamic content.

Figure 1 Highlighting different kinds of changes on a website using the VDiff algorithm for differencing HTML content. Source: [2]

Analysis of the dynamics of web content can also be based on tracking terms and their use accross websites [1]. Terms can define a temporal content, that can appear on websites for a short time, for example based on current news. They can also describe a seasonal content, that appears periodically in connection with some other repeating events. Observing the use of terms on websites can provide a basis for making connections between them and detecting the flow of information on the web.

The flow of information can also be observed in sharing multimedia content, such as videos or photos. These connections can be represented in a graph to show the flow of content accross the web.

In our work, we aim to design and implement a method to effectively process Web content in order to be able to observe and analyse its behaviour in an archived data set. We plan to use the method on a sufficiently large data set of websites from various sources (e.g. blogs, social networks or news portals) to draw useful conclusions about the dynamics of such content.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas (2009): The web changes everything: understanding the dynamics of web content. Web Search and Data Mining 2009, ACM, pp. 282–291.

[2] Rimon Mikhaiel and Eleni Stroulia (2005 Accurate and Efficient HTML Differencing. In Proceedings of the 13th IEEE International Workshop on Software Technology and Engineering Practice (STEP '05).