

Extraction of Contextual Metadata from File System Interactions

Adaora Okoli¹ and Bernhard Schandl²

¹ Smart Information Systems, Vienna, Austria, a.okoli@smart-infosys.net

² University of Vienna, bernhard.schandl@univie.ac.at

Semantic systems improve information management by providing mechanisms to access objects by their attached semantic metadata elements. Semantic file systems apply this principle to file systems and aim to provide more user orientation in file management. Rather than identifying files by their fixed name and position in a strict hierarchy they can be addressed by their semantics, which are attached to files in the form of explicit metadata tags, structured attributes, controlled annotation vocabularies, or semantic links.

However, in order to be able to benefit from the inherent capabilities of semantic file systems, users are often required to manually enrich their files with meaningful and useful metadata, which is an expensive and complex task users might not always be willing to perform. Thus, to untie the basic usefulness of semantic file systems from explicit user participation, we propose to automatically generate metadata by analyzing file interaction sequences, i.e., tracing read and write access operations to file data.

The basis for the automatic generation of contextual relationships is the definition and detection of contexts, and the identification of files that are relevant in the same contexts. It can be supposed that the temporal locality of file interactions corresponds to the relatedness of the accessed files [Har06]; thus, it is the least obtrusive manner to approach the automatic generation of contextual file relationships by analyzing time-ordered file interaction sequences and deriving higher-level semantic annotations from these relationships.

We have implemented a virtual file system [SH09] that represents all file metadata as RDF triples. It allows for semantic annotation and retrieval through a dedicated query API or through a standards-compliant SPARQL endpoint. This infrastructure allows us to capture all user interactions on files within the virtual file system and to perform analysis algorithms on it. The results of the analysis algorithms are then written back as file annotations and hence can directly support the user in their information retrieval tasks.

References

- [Har06] Chris Harrison. Kronosphere - A temporal visualization for file access. Master's thesis, Department for Computer Science, New York University, USA, May 2006.
- [SH09] Bernhard Schandl and Bernhard Haslhofer. The Sile Model – A Semantic File System Infrastructure for the Desktop. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, Heraklion, Greece, 2009.