Living Hypertext — Web Retrieval Techniques

Ralf-Dieter Schimkat\textsuperscript{1}, Wolfgang Küchlin\textsuperscript{1}, and Frank Nestel\textsuperscript{2}

\textsuperscript{1} Wilhelm-Schickard Institute for Computer Science
\url{http://www-sr.informatik.uni-tuebingen.de}
Sand 13, 72074 Tübingen, Germany
\{schimkat,kuechlin\}@informatik.uni-tuebingen.de
\textsuperscript{2} INA-Schaeffler KG
Industriestrasse 1–3, 91074 Herzogenaurach, Germany
Frank.Nestel@de.ina.com

Abstract. In this paper, we present a document metaphor called Living Documents for accessing and searching for digital documents in modern distributed information systems. Our approach is based upon a fine-grained document concept which glues computational services, data and meta data together. Viewing documents as micro servers is particularly well suited in environments where the document’s content is changing continuously and frequently. Based on a case study of an existing state-of-the-art Web application, we show how to transform database-centric information systems into a hypertext of inter-linked Living Documents. We also discuss how to effectively use traditional as well as Web information retrieval techniques, namely topic distillation, in such hypertext environment. In particular, an extended version of Kleinberg’s \cite{11} algorithm is presented.

1 Introduction

It is generally agreed upon that the major task in information retrieval is to find relevant documents for a given query \cite{9}. A document is a collection of digital information ranging from plain text files, data-related meta attributes to multi-media documents. Clearly a conceptual model of information retrieval dealing with text or multi-media documents should integrate different views on documents. Agosti et al. \cite{11} point out that one of several differences in traditional information retrieval and information retrieval on the Web is the different kind of management of the collection of documents. In fact the web is a virtual collection since a real collection stored at one particular location such as one single database would be unmanageable.

In contrast to the Web, traditional information system design is about documents which are made persistent in a digital document archive, but their attributes (meta data) are kept in databases. Even virtual documents missing any digital content can be seen as an aggregation of their attributes. This leads to several potential drawbacks in classically designed information systems: (i) the specification of document attributes is bound to the database schema and is a priori determined at the time the database schema is set up. (ii) Furthermore,
the static specification restricts a document’s life cycle. For example, in most cases it is hard to add new kinds of meta data to the document’s collection at run time.

Our goal is to provide a document metaphor and an implementation respectively which can be used in traditional as well as in Web information systems. Our approach is characterized by transforming documents into active containers managing their content and meta data in an uniform and extensible manner.

The main contributions of this paper are: (i) Introduction to a new document metaphor called Living Documents. (ii) Description of a complete implementation path of Living Documents based on a case study of a contemporary web information system. We centered our implementation around the concepts of mobile agents and general data description languages based on XML. (iii) We show how to deploy three different kinds of information retrieval techniques in Living Documents. In particular, we describe how to use state-of-the-art web information retrieval techniques as topic distillation within a web of inter-linked Living Documents. Finally, we present an extension of a well-known algorithm for the analysis of connectivity graphs in hypertext environments.

2 Living Documents

First, we give an introduction to the concept of Living Documents from an abstract point of view neglecting any implementation details. In the next section we show an implementation path for Living Documents.

![Diagram of Living Documents](http://www.living-documents.org)

Fig. 1. A) Components of a Living Document. A Living Document is divided into three sections: Raw Data carries the document to manage, Semi-Structured Data contains all meta data about the managed documents, and the Code section keeps the computational services for accessing a LD and processing incoming requests (i.e. queries). B) Digital Documents are turned into Living Documents. Living Documents form a hypertext by keeping links similar to hypertext links to each other.

---

3 Home page of Living Documents at [http://www.living-documents.org](http://www.living-documents.org)