Semantic Document Networks to Support Concept Retrieval

S

Simon Boese
University of Hamburg, Germany

Torsten Reiners
Curtin University, Australia & University of Hamburg, Germany

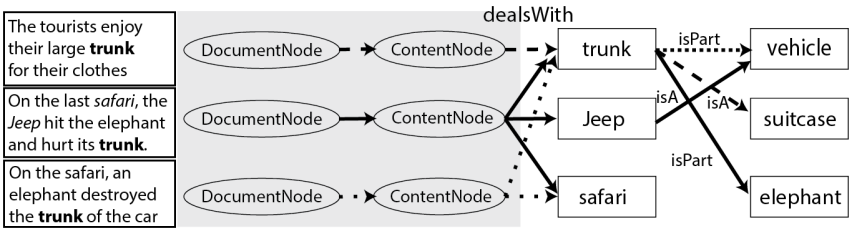
Lincoln C. Wood
Auckland University of Technology, New Zealand & Curtin University, Australia

INTRODUCTION

This chapter focuses on a framework to support advanced document storage and fast queries to retrieve documents based on concept-focused searches. These searches favour 'semantic' searches which evaluate and use the meanings of words and phrases, rather than 'key-word' searches. The framework rests on three stages: *pre-processing* (semantic analysis influences the storage quality within a semantic database), *conceptualization* (extraction of key concepts to establish document networks), and *storage* within a semantic database, facilitating advanced future retrieval. The objective is to decompose documents and extract all relevant information about structure and content to allow comprehensive storage in a semantic document network; including the interpretation according to domains, contexts, languages, or readers. For example, the word 'trunk' may refer to a storage area (in the context of motor vehicles), a clothes storage box (in the context of travelling), or an elephant's appendage (in the context of a safari); see Figure 1. The arrows represent parameters associated with relations. There can be multiple meanings for the related words and it is only the clustering of words that provides the important context which provides readers with meaning; e.g., Safari is also the name of an Internet browser.

A brief introduction to conceptualization and the semantic document network provides an overview of how information can be stored in an interlinked network. Using a short sample, we demonstrate the calculation of the semantic core using concept-based indexing and how the concepts are embedded within the existing semantic document network.

Figure 1. Evaluation of the meaning of 'trunk' based on the context. This supports semantic-based retrieval of documents rather than merely keyword-based retrieval [Source: Boese, Reiners, and Wood (2012, p. 5)].



DOI: 10.4018/978-1-4666-5202-6.ch192

BACKGROUND

Organizations are facing increasingly significant document management challenges as they seek to leverage vast volumes of internally-focused documents (e.g., emails or internal reports) or provide document-based services to others. The challenge is to design document management systems that support the storage and retrieval of *unstructured* electronic documents; in contrast, there are well-established document management methods for *structured* documents, such as those used by libraries. Limited meta-information (particularly key terms) has historically been used to support simple indexing and classification procedures. However, the rise of user-generated content within Web 2.0, and the on-going accumulation of document digitalization have led to the challenge to maintain, let alone increase, the retrieval quality. Improved search engine capabilities enable users to consider synonyms, stem forms, and even translations (He & Wang, 2009). However, these elements share the commonality of requiring a search request that is based on words within the document, while ignoring the meaning and context that these words occur in – they ignore the semantic meaning behind the text. Semantic analysis can support the search through the determination of the key concepts and scenarios that may be associated with a term; e.g., the word 'trunk' may be used with a different meaning in documents about car repair, travel accessories, or in safari reports. As the Web progresses and evolves, we anticipate that computers will continue to process information on increasingly higher levels, and will soon enable search and retrieval of documents based on the meaning of words, rather than just the occurrence of words. The underlying systems that support this process would also enable other applications for handling documents, enabling software agents to extract individualised information from databases, grade unstructured exams with minimal instructor setup, summarise correspondences or articles, and translate documents effectively. In all of these cases, the ability to understand natural, unstructured language is crucial to ensure the robustness and reliability of the results.

CONCEPT RETRIEVAL WITH SEMANTIC DOCUMENT NETWORKS

A *concept* is described by one or multiple words and is associated with a (semantic) category. These categories represent the meanings or senses of the containing words, whereas the interpretation might differ between domains, contexts, languages, or readers (Davies, 2009). Conceptualization is the process of detecting texts' meaning as provided by a set of connected concepts. The goal is to identify descriptive, yet generic terms that characterise the entire text. Such concepts reduce the text to its most relevant elements with respect to the textual content and can be regarded as a 'footprint'. Two different texts could be seen as identical, in terms of the story they hold or the message they are delivering, if their footprint, composed of concepts linked together, matches.

Humans intuitively employ their cognitive abilities to undertake conceptualization, which facilitates the generalization or abstraction from the full text. Working with the premise that the reader is familiar with the vocabulary, a reader can understand the text and deduce knowledge from the document (Reiterer, Dreher, & Gütl, 2010). The challenge is to develop software that is capable of replicating this process. This conceptualization must also cover various scenarios, such as when a writer may discuss the 'trunk' (of the elephant they photographed on holiday) while in a garage, next to an automobile (which also has a 'trunk' in the rear of the vehicle); such juxtapositions result in an erroneous footprint, leading to later misinterpretation by many automated software approaches.

The process of conceptualization has many applications. A key reason for use is that it can provide a deduction of a 'concept hierarchy', allowing a user to trace from an abstract concept to a more concrete concept, or vice versa. This