

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, unstruc-

tured 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

10 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/semantic-document-networks-to-support-concept-retrieval/107400

Related Content

BISC: A Framework for Aligning Business Intelligence with Corporate Strategies Based on Enterprise Architecture Framework

Atieh Dokhanchi and Eslam Nazemi (2016). *Business Intelligence: Concepts, Methodologies, Tools, and Applications* (pp. 501-518).

www.irma-international.org/chapter/bisc/142635

Contracts and Contracting Styles: Existing and Future Types

Mohamed Bassyouni (2021). *Innovative and Agile Contracting for Digital Transformation and Industry 4.0* (pp. 18-39).

www.irma-international.org/chapter/contracts-and-contracting-styles/272632

A Systematic Approach for Business Data Analytics with a Real Case Study

Kaibo Liu and Jianjun Shi (2015). *International Journal of Business Analytics* (pp. 23-44).

www.irma-international.org/article/a-systematic-approach-for-business-data-analytics-with-a-real-case-study/132800

Data-Driven Decision Making for New Drugs: A Collaborative Learning Experience

George P. Sillup, Ronald K. Klimberg and David P. McSweeney (2010). *International Journal of Business Intelligence Research* (pp. 42-59).

www.irma-international.org/article/data-driven-decision-making-new/43681

Monitoring Time Consumption in Complementary Diagnostic and Therapeutic Procedure Requests

Ana Alpuim, Marisa Esteves, Sónia Pereira and Manuel Filipe Santos (2016). *Applying Business Intelligence to Clinical and Healthcare Organizations* (pp. 208-240).

www.irma-international.org/chapter/monitoring-time-consumption-in-complementary-diagnostic-and-therapeutic-procedure-requests/146070