Encyclopedia of Business Analytics and Optimization

John Wang
Montclair State University, USA

Volume IV
Op-So

BUSINESS SCIENCE Reference
An Imprint of IGI Global

Managing Director: Lindsay Johnston
Production Editor: Jennifer Yoder
Development Editor: Austin DeMarco
Acquisitions Editor: Kayla Wolfe
Typesetter: Christina Barkanic, Michael Brehm, John Crodian, Lisandro Gonzalez, Christina Henning, Deanna Jo Zombro
Cover Design: Jason Mull

Published in the United States of America by
Business Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of business analytics and optimization / John Wang, editor.
pages cm
Includes bibliographical references and index.
Summary: "This reference confronts the challenges of information retrieval in the age of Big Data by exploring recent advances in the areas of knowledge management, data visualization, interdisciplinary communication, and others"-- Provided by publisher.
ISBN 978-1-4666-5202-6 (hardcover) -- ISBN 978-1-4666-5203-3 (ebook) -- ISBN 978-1-4666-5205-7 (print & perpetual access) 1. Management--Mathematical models. 2. Decision making--Mathematical models. 3. Business planning--Mathematical models. 4. Big data. I. Wang, John, 1955-
HD30.25.E53 2014
658.4'038--dc23
2013046204

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Semantic Document Networks to Support Concept Retrieval

S

Simon Boese
University of Hamburg, Germany

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

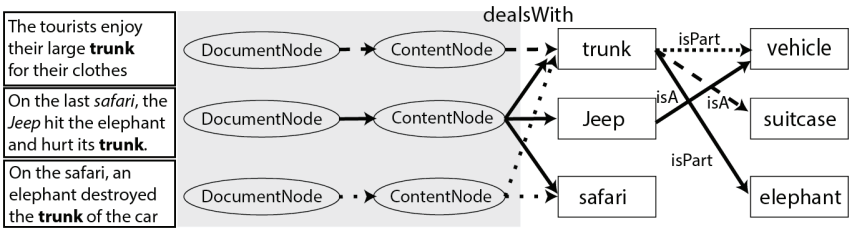
Lincoln C. Wood
Auckland University of Technology, New Zealand & Curtin University, Australia

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

process is important in *information retrieval*, meaningful retrieval of stored information from a data repository with digital artefacts (Micarelli, Sciarrone, & Marinilli, 2007); *automated essay grading*, to compare essays using their 'feature space' of relevant concepts and their interrelationships and comparing these to model answers (Toranj & Ansari, 2012; Dikli, 2006); *detection of plagiarism*, identifying reproduced ideas and structuring of concepts as opposed to identifying perfectly copied segments of text (Osman, Salim, Binwahlan, Alteeb, & Abuobieda, 2012); and *ontology construction*, where an ontology is formed from text relations and both generalization and specialization of concepts (Buitelaar, Cimiano, & Magnini, 2005), using the complex interplay between these relations to construct an ontology schema.

Technology-supported identification of concepts is a growing area of interest and multiple technologies are required. A complete explication is beyond our scope (see Allen (1995) for further details), yet four core technologies are:

- **Tagging Part Of Speech (POS):** Is the classification of string sequences, separated by white spaces, annotated with the determined grammatical construct (e.g., noun or verb). Punctuation provides interpretation of the structure and the ability to identify subordinate clauses (Allen, 1995).
- **Named Entity Recognition (NER):** Identifies different uses of terms or expressions for the same, known, objects, people, or places (Tjong Kim Sang & De Meulder, 2003). Detection relies on the coverage of underlying entity lists.
- **Entity Tracking (ET):** Identifies objects referenced differently over a document (Maynard, Bontcheva, & Cunningham, 2003); a problem known as 'co-referencing'. They may be synonyms, named entities, or pronouns ("Mr West vanished. He hasn't been since.") (Florian, Hassan, Ittycheriah, Jing, Kambhatla, Luo, Nicolov, & Roukos, 2004).
- **Relation Extraction (RE):** Infers relationships between verbs and involved objects, exploiting grammatical structures. Active or passive roles are assigned to participating objects, with relations often expressed as predicates (Harabagiu, Bejan, & Morărescu, 2005; Koenig, Mauner, Bienvenue, & Conklin, 2008).

Methodologies Often Employed

Working on the premise that most texts are unstructured or semi-structured, different techniques have been developed to retrieve relevant concepts from documents.

- **Formal Concept Analysis (FCA):** Is a mathematical approach to deduce knowledge from texts, using lattice theory, assuming partially ordered sets (Wille, 2009). Fundamental to FCA is the definition of 'context' as expressed in a tuple, a set of entities or attributes, and their binary relationships (Ganter, Stumme, & Wille, 2005; Ferré & Rudolph, 2009; Wille, 2009). A subset defines a concept and concept hierarchies that can be processed using set notation (Carpineto & Romano, 2004).
- **Latent Semantic Analysis (LSA):** Infers contextual relations between terms (Landauer, 2007; Martin et al., 2007), working on the premise that the relationships are hidden and must be inferred (Landauer, 2007). A correlation matrix of terms and occurrences is constructed, recording the contexts in which terms occur. Each row is considered as an independent group of terms. A weighting function is applied to transform the matrix, before it is split into three separate matrices using the process of Singular Value Decomposition (SVD). Dimensional reduction of the SVD, employing the original correlation matrix, reduces the noise and enables identification of the most important terms, or concepts.

- **Concept-based Indexing (CBI):** First assembles a semantic core that is stored within a semantic network, in which nodes represent important concepts identified on the basis of a semantic relatedness. The result can describe the content of documents, contributing directly to the creation of a semantic document network. This is the methodology that we discuss further in the following section.

Concept-Based Indexing

First, appropriate concept candidates need to be identified. These are adjacent, compound terms, or single words. For multiword concepts the *synsets*, sets of semantically equivalent terms, can be retrieved from specialised databases like WordNet. WordNet is a Princeton University-developed lexical database that stores concepts within synsets comprising of synonyms and gloss, or definition. A single given word may belong to multiple synsets, reflecting the ambiguity inherent in the language. Concept candidates must be weighted, where they receive a higher weight when they appear more frequently across all documents that are being analysed in a single pass. This overall weighting is then compared to the current document with a final score being calculated.

For each concept, the number of times that it occurs in the text, and the number of words it contains, is recorded. This is assembled within a particular set which holds all sub concepts. All of the elements of partially ordered subsets of the words are included; e.g., "House of Commons" includes each separate word as well as the terms *House of* and *of Commons*. Concept candidates that have a weight exceeding an established threshold will be retained. This enables to keep the focus on the core, crucial concepts.

Second, the semantic relatedness of these concept candidates will be calculated using an adapted *Lesk algorithm* (Lesk, 1986). This algorithm resolves ambiguities between words using a machine readable dictionary and overlapping words definitions are detected (Banerjee & Pedersen, 2010; Lesk, 1986). Multiple WordNet relations over synsets are used to compute the overlap between concept candidates within the set. The overlap is computed for all meanings of the candidate concept, in combination with meanings of all other concept candidates, forming the given set deduced from WordNet using FCA (Miller, 1998). This overlap is the *Lesk Overlap*, defined as the sum of conjoined words, where the number of adjacent, conjoined words is squared (Baziz, Boughanem, & Traboulsi, 2005; Koenig et al., 2008).

The semantic core is then determined for all the possible combinations of concept meanings. For each of the potential meanings of concept candidate the relatedness to all the other candidates' meanings are aggregated, forming an overall *meanings score* for the concept candidate. The most highly scoring meaning for each concept candidate is selected to compose a 'semantic core' and is accepted as the actual meaning in that document. These calculated concept meanings represent the nodes within the semantic core. Edges between the nodes represent the semantic relatedness of the corresponding meanings.

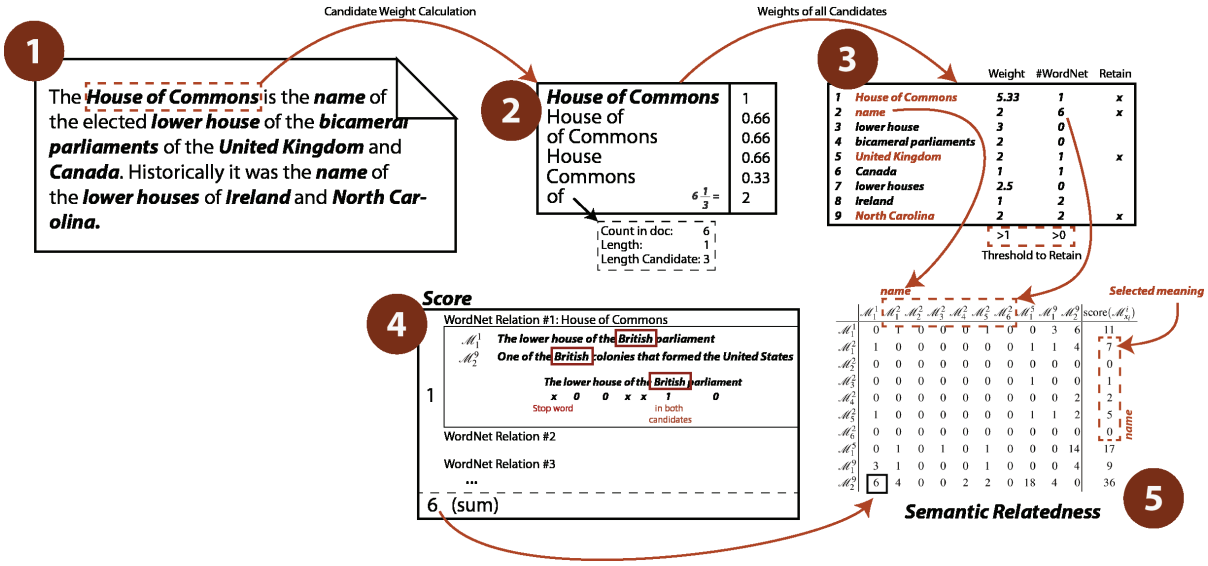
An Example of Concept-Based Indexing

An exemplary segment of text is used to demonstrate the algorithm. Within the text shown below are the identified concept candidates that have been highlighted (Boese, 2012):

*"The **House of Commons** is the **name** of the elected **lower house** of the **bicameral parliaments** of the **United Kingdom** and **Canada**. Historically it was the **name** of the **lower houses** of **Ireland** and **North Carolina**"*

These candidates are selected by checking each word and determining whether it belongs to a defined 'stop word' category (viz., articles, prepositions, verbal forms and types, pronouns,

Figure 2. An overview of the process used to evaluate the semantic relatedness of concepts, a crucial sequence of steps in the development of the semantic core

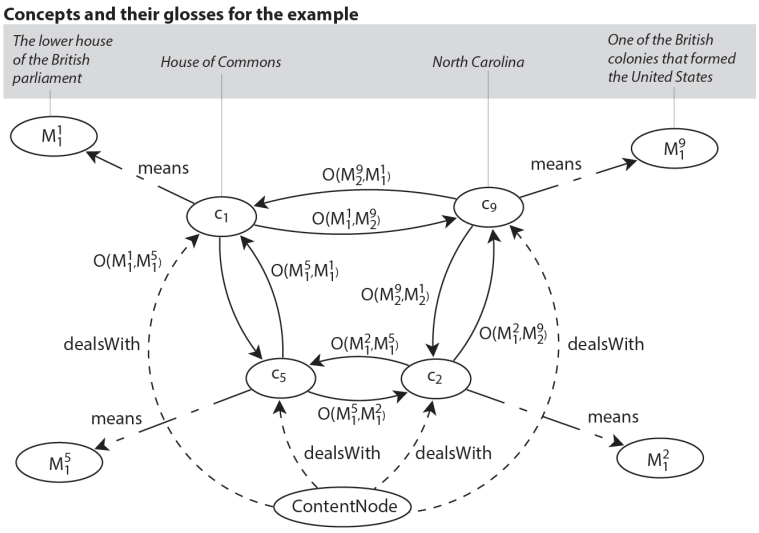


and verbs, symbols, list markers, and conjunctional words). Exceptions are made for entities where a stop word is part of the candidate; e.g., "House **of** Commons." The identified terms are retained as concept candidates. The further steps of the algorithm are visualised in Figure 2.

1. Identify concept candidates; see above.
2. The weight of each candidate is computed. The total weight is calculated by sum the weights of all subsets of the candidates' words; here *house of commons, House of, of Commons, House, Commons*, and *of*. The weight is the relation of the candidates' length (number of words; here 3 for *House of Commons*) and the specific subset (e.g., 1 for *of*) multiplied by the total occurrence in the document (for the given text it is 6 times). In this example, the weight for *House of Commons* is 5.33.
3. Only concept candidates with a weight above a given threshold are retained for the semantic core calculation. In this example, the threshold is 2, eliminating the candidates 6 and 8. Candidates 3, 4, and 7 are not retained despite their weight as they do not have entries in WordNet that was used as a lexical database.
4. The semantic relatedness of all the concept candidates' meanings is then calculated. The example depicts this procedure for the candidate "House of Commons" and one meaning of North Carolina. The score is calculated by overlapping concepts excluding stop words; resulting in a score of 1 for the common concept *British*. Note that the total score for the relatedness of *House of Commons* and *North Carolina* is 6 if the other meanings are compared to each other.
5. The score in the right column shows the overall score of each concept meaning calculated by summing all columns of the specific row. In case of multiple WordNet entries, the one with the best score is chosen for the semantic core. In this example, the candidate concept *name* had 6 WordNet entries with 6 corresponding scores; whereas the first one (M^2_1) has the highest score with 7 and, therefore, is chosen for the semantic core.

Figure 3 illustrates a simplified semantic core. The concepts c1 (*House of Commons*), c2 (*name*),

Figure 3. Concepts and their glosses for the "House of Commons" example [Source: Boese et al. (2012, p. 5)]. The notation in the figure is: (M^x_y) with x being the candidate number in the core (see Step 3 in Figure 2), y being the number of meaning for this candidate extracted from WordNet (see Step 4 in Figure 2). C_x is the concept x.



c5 (*United Kingdom*), and c9 (*North Carolina*) as well as their calculated best meanings based on the lexical database WordNet form the semantic core that best represents the document. The concept nodes are connected to a *Content Node* for integration into the document network.

The Semantic Document Network

The transition of documents into a semantic network is generally restricted to the meaning of the documents, rather than the included structure itself. The preparation for automated pre-processing of documents is relevant where the derived implications are based on existing knowledge or reasoning using given rules (Petöfi, 1976).

The Partial Networks

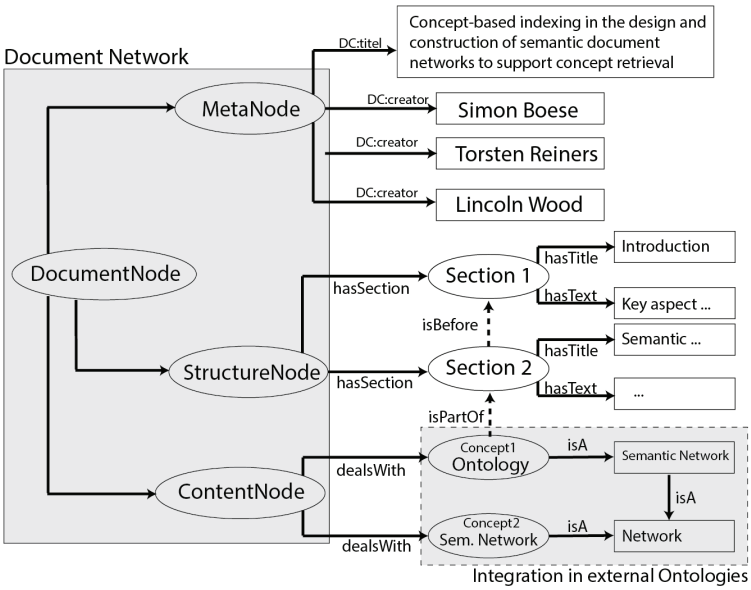
Content representation is only one necessary component. The preservation of documents within the semantic networks requires the encoding and storage of document structures and corresponding meta-information. Petöfi (1976) suggests that a document node should connect different networks with dimensions that must be related. The partial networks are disjunct, promoting independent parallel constructions. Nevertheless, these individual nodes from one network may be connected to others via links; e.g., a concept relates to a specific section. We distinguish between information nodes (representing structure, the oval elements in Figure 4) and content nodes (representing literals, the square elements in Figure 4).

The meta-information network contains all of the information used to describe the document itself; e.g., authors and date of publication. The commonly used standard is the *Dublin Core*, which defines annotations for a relation between two elements in the document or digital artefact.

The network structure defines a logical structure of a document. This describes the order of text units (e.g., chapters) and the content (e.g., text or images). Links between the nodes of the structure described dependencies such as order, or nested units.

The network concept represents the content of the document structured by concepts. Extracted

Figure 4. An exemplary semantic document network that visually connects the components [Source: Boese et al. (2012, p. 5)]. 5].



concepts are connected amongst one another and also and externally defined ontologies, providing additional information regarding superior- or subordinate-concepts. This allows a comprehensive analysis of the document (Schubert, 1991; Shastri, 1991). Concepts can have three states: *fully-specified*, where they are known and form part of an ontology; *partly-specified*, where further information such as concepts within the partial network are required to fully evaluate concept; and *under-specified*, where they are unknown in this context.

These partial networks are constructed independently; however, links between the structure and content network encode valuable information. Concepts are allocated to specific structural units and define the order within the concepts.

Network Consistency

Document networks must be validated against consistency constraints prior to being added to a repository. This step guarantees the overall quality within the repository and ensures complete document information is maintained. Three key constraints are: 1) the existence of partial networks, 2) exceeding a minimum amount of information within partial networks, and 3) valid implications for relations and inheritance of literals.

1. A document network can only be complete if it has two component partial networks: the *content* and *meta-information* networks. These networks are mutually dependent; concepts within the content require a context, specified within the meta-information. However, if the meta-information is obsolete then there is no additional information about the content itself. The structure network is optional as not all documents will have recognizable structure; e.g., short documents.
2. In addition to partial networks, it is mandatory to have elements in the network that identify the document and allow it to be added on to the semantic document network. This will depend on the type of document being analysed; if the structure network is provided, this must contain adequate information to reconstruct the structure of the document. For simple documents such

meta-information may include only the title, author, and one concept that is part of an ontology for classification.

3. The semantic network defines the hierarchical relationships between components of the structure; e.g., paragraph 1.1 is part of paragraph 1. The same applies to concepts if they are associated to an ontology; the relation between the two concepts must be valid when accounting for the ontology hierarchical structure. The given system must be able to perform semantic verification and detect inconsistencies while processing the document.

Semantic Database

Following the construction of the semantic core and the document network, these elements need to be stored within the database. The use of a specialised semantic document network rather than relational databases improves retrieval and prevents overlap with other documents. Within this repository documents share common nodes; i.e., they have similarity and closeness. To prevent association of terms to the wrong document, disambiguation methods are used where all relations are parameterized with the document identification, allowing for later reconstruction of individual document networks. Examples of semantic databases include Neo4J, OWLIM; see also Aust and Sarnow (2009).

Mode of Operation

A data model of a semantic network database represents information using two elements: nodes and edges. Nodes constitute atomic, unique values which imply that each node only exists once within the database. Labelled edges are an association between two nodes and embody direct relations between the nodes.

Some semantic network databases allow only additions and deletions of nodes and do not support updates on existing nodes. This ensures consistency; all connections must be removed explicitly, ensuring relationships remain current. Modification of nodes and edges is accomplished through the deletion and then insertion of new elements. Design variants should be able to support nodes and edges that may be under-, partly-, or fully-specified. This is useful when an exact value is unknown, but it is known that a value does exist (i.e., it is under-specified) and when an exact value that fills a pattern that may be expressed as a logical expression (i.e., it is partly-specified).

Data types are generally assigned. Within some semantic network databases values may not be assigned directly to certain types. Additional information is required for typification; i.e., at least one node and connecting edge to describe the relationship. Nodes may implicitly be an instance of different types as there is no mechanism that restricts the number of edges on a single node; the situation necessitates a way to guarantee particular properties. Constraints can be specified as semantic networks themselves, and utilise different specified network elements to ascertain the existence of other nodes and edges. One or more edges are declared as a trigger, defining the responsibility of a constraint network. When a trigger 'fires' the constraints claim the other elements and reject the network.

The Database Layer

The completed semantic document network is validated against a constraint network to ensure consistency. If the required network elements are not all present, the document network will be rejected and conceptualization will be reattempted using different parameters. All elements of the document network will be stored within the database, sending the entire set of edges and nodes into the semantic database. Each edge is aware of the start and end node ensuring that no nodes remitted during this process.

FUTURE RESEARCH DIRECTIONS

It is clear that improvement in semantic document network storage is an area that requires further research. However, the integration of these concepts in other domains such as automated essay grading or the evaluation of machine translations is also important. This Concept-Based Indexing framework enables semantic document networks to be incorporated within other domains where existing systems can take advantage of semantic pre-processing. Further work can also be undertaken to allow automated mapping of business processes, constructed by the examination of existing corporate documentation, which could generate updates to existing business processes and provide alerts where processes are not being adhered to. This may lead to improved flexibility, auditability, and adaptability within supply chains.

CONCLUSION

Future information and communication technologies will enable the interpretation of unstructured documents to extract meaning. The growth of the Web 2.0, allowing users to create content, has generated an exponential proliferation of documents beyond the ability of humans to process. Therefore, it is important to develop intelligent systems that may automate the processing, extraction, and retrieval of information, enabling other users and systems to use the collected documents.

We have presented a framework with focus on the storage of processed documents, covering:

- the determination of best concept candidates using concept based indexing,
- the definition of semantic cores for documents,
- the generation of a semantic document network were each document is comprised of the partial networks *meta-information*, *structure*, and *concept*, and
- the storage process for documents into semantic network databases.

REFERENCES

Allen, J. F. (1995). *Natural language understanding*. Redwood City, CA: Benjamin/Cummings.

Aust, T., & Sarnow, M. (2009). *Entwurf und implementierung einer medien-datenbank-middleware mit integrierten semantischen netzen [English: Design and implementation of a media database middleware using semantic networks]* (Diploma Thesis). University of Hamburg, Germany.

Banerjee, S., & Pedersen, T. (2010). An adapted Lesk algorithm for word sense disambiguation using WordNet. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (Vol. 2276, pp. 136–145). Lecture notes in computer science Berlin, Heidelberg, Germany: Springer. doi:10.1007/3-540-45715-1_11

Baziz, M., Boughanem, M., & Traboulsi, S. (2005). A concept-based approach for indexing documents in IR. *Actes du XXIIIème Congrès INFORSID* (pp. 489-504).

Boese, S., Reiners, T., & Wood, L. C. (2012). *Design and construction of semantic document networks using concept extraction*, School of Information Systems Working Paper Series, Curtin University of Technology, Australia.

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). *Ontology learning from text: Methods, evaluation and applications* (1st ed.). Amsterdam: IOS Press.

Carpineto, C., & Romano, G. (2004). *Concept data analysis* (1st ed.). Sussex, England: Wiley. doi:10.1002/0470011297

Davies, J. (2009). *Semantic knowledge management*. Berlin, Germany: Springer. doi:10.1007/978-3-540-88845-1

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1), 1–35.

(2009). InFerré, S., & Rudolph, S. (Eds.). Lecture notes in artificial intelligence: Vol. 5548. *Formal concept analysis*. Berlin, Heidelberg, Germany: Springer. doi:10.1007/978-3-642-01815-2

Florian, R., Hassan, H., Ittycheriah, A., Jing, H., Kambhatla, N., & Luo, X. et al. (2004). *A statistical model for multilingual entity detection and tracking*. Yorktown Heights, NY: I.B.M. T.J. Watson Research Center.

Ganter, B., Stumme, G., & Wille, R. (2005). *Formal concept analysis: Foundations and application*. Berlin, Germany: Springer.

Harabagiu, A., Bejan, C. A., & Morărescu, P. (2005). Shallow semantics for relation extraction. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '05)* (pp. 1061–1066).

He, D., & Wang, J. (2009). Cross-language information retrieval. In A. Göker, & J. Davies (Eds.), *Information Retrieval: Searching in the 21st Century* (pp. 234–254). Chichester: Wiley. doi:10.1002/9780470033647.ch11

Koenig, J.-P., Mauner, G., Bienvenue, B., & Conklin, K. (2008). What with? The anatomy of a (proto)-role. *Journal of Semantics*, *25*(2), 175–220. doi:10.1093/jos/ffm013

Landauer, T. K. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *SIGDOC '86: Proceedings of the Annual International Conference on Systems Documentation* (pp. 24–26). New York, NY: ACM Press. doi:http://doi.acm.org/10.1145/318723.318728

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind latent semantic analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Knitsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–55). Mahwah, NJ: Lawrence Erlbaum Associates.

Micarelli, A., Sciarrone, F., & Marinilli, M. (2007). Web document modeling. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive Web: Methods and strategies of Web personalization* (Vol. 4321, pp. 155–192). Lecture Notes in Computer Science Berlin, Germany: Springer. doi:10.1007/978-3-540-72079-9_5

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.), *WordNet: An electronical lexical database* (pp. 25–46). Cambridge, MA: MIT Press.

Osman, A. H., Salim, N., Binwahlan, M. S., Alteeb, R., & Abuobieda, A. (2012). An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing*, *12*(5), 1493–1502. doi:10.1016/j.asoc.2011.12.021

Petöfi, J. S. (1976). *Text representation and lexicon as semantic network. Logiche, calcoli, formalizzazioni e lingue storico-naturali*. Catania: Italien.

Reiterer, E., Dreher, H., & Gütl, C. (2010). Automatic Concept Retrieval with Rubrico. In M. Schumann, L. M. Kolbe, M. H. Bretiner, & A. Frerichs (Eds.), *Anwendung der konzeptanalyse und ontologische modellierung in der wirtschaftsinformatik, (MKWI 2010)* (pp. 3–14). Universitätsverlag Göttingen.

Schubert, L. K. (1991). Semantic nets are in the eye of the beholder. In J. F. Sowa (Ed.), *Principles of semantic networks: Explorations in the representation of knowledge, The Morgan Kaufmann Series in Representation and Reasoning* (pp. 95–108). San Mateo, CA: Kuafmann.

Shastri, L. (1991). Why semantic networks? In J. F. Sowa (Ed.), *Principles of semantic networks: Explorations in the representation of knowledge, The Morgan Kaufmann Series in Representation and Reasoning* (pp. 109–136). San Mateo, CA: Kaufmann.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In M. Daelemans, & O. Walter (Eds.), *Proceedings of the conference on natural language learning at HLT-NAACL 2003* (pp. 142–147). Morristown, NJ: Association for Computational Linguistics.

Toranj, S., & Ansari, D. N. (2012). Automated versus human essay scoring: A comparative study. *Theory and Practice in Language Studies*, *2*(4), 719–725.

Wille, R. (2009). Restructuring Lattice Theory: An approach based on hierarchies of concepts. In S. Ferré, & S. Rudolph (Eds.), Lecture notes in artificial intelligence: Vol. 5548. Formal concept analysis (pp. 314–339). Berlin, Heidelberg, Germany: Springer.

ADDITIONAL READING

Baruzzo, A., Casoto, P., Challapalli, P., Dattolo, A., Pudota, N., & Tasso, C. (2009). Toward semantic digital libraries: Exploiting Web2.0 and semantic services in cultural heritage. *Journal of Digital Information*, *10*(6). Retrieved from http://journals.tdl.org/jodi/index.php/jodi/article/view/688/576

Boonthum-Denecke, C., McCarthy, P. M., & Lamkin, T. A. (2011). *Cross-disciplinary advances in applied natural language processing: Issues and approaches*. Hershey: IGI Global. doi:10.4018/978-1-61350-447-5

Dong, H., Hussain, F., & Chang, E. (2008). A survey in semantic search technologies. In *2nd IEEE International Conference on Digital Ecosystems and Technologies* (pp. 403–408).

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. [TOIS]. *ACM Transactions on Information Systems*, *29*(2), 8. doi:10.1145/1961209.1961211

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., & Motta, E. (2011). Semantically enhanced information retrieval: An ontology-based approach. *Web Semantics: Science. Services and Agents on the World Wide Web*, *9*(4), 434–452. doi:10.1016/j.websem.2010.11.003

Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive development* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Göker, A., & Davies, J. (Eds.). (2009). *Information retrieval: Searching in the 21st century*. Chichester, UK: Wiley. doi:10.1002/9780470033647

Juršič, M., Sluban, B., Cestnik, B., Grčar, M., & Lavrač, N. (2012). Bridging concept identification for constructing information networks from text documents. In M. R. Berthold (Ed.), *Bisociative knowledge discovery* (Vol. 7250, pp. 66–90). Lecture notes in computer science Berlin, Heidelberg, Germany: Springer. doi:10.1007/978-3-642-31830-6_6

Kruk, S., & McDaniel, B. (2009). *Semantic digital libraries*. Berlin, Germany: Springer. doi:10.1007/978-3-540-85434-0

Odijk, D., de Rooij, O., Peetz, M. H., Pieters, T., de Rijke, M., & Snelders, S. (2012). Semantic document selection theory and practice of digital libraries. In P. Zaphiris, G. Buchanan, E. Rasmussen, & F. Loizides (Eds.), *Theory and practice of digital libraries* (Vol. 7489, pp. 215–221). Lecture notes in computer science Berlin, Heidelberg, Germany: Springer. doi:10.1007/978-3-642-33290-6_24

Song, M., & Wu, Y.-F. B. (2009). *Handbook of research on text and Web mining technologies*. Hershey, PA: IGI Global.

Staab, S., & Studer, R. (2009). *Handbook on ontologies*. Berlin, Germany: Springer. doi:10.1007/978-3-540-92673-3

Sugumaran, V. (2011). *Applied semantic Web technologies*. Boca Raton, FL: CRC Press.

Van Atteveldt, W. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media Content*. Charleston: BookSurge.

Van Britsom, D., Bronselaer, A., & De Tré, G. (2012). Concept identification in constructing multi-document summarizations. In S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, R. R. Yager (Eds.), Communications in computer and information science: Vol. 298. Advances in computational intelligence (pp. 276–284). Berlin, Germany: Springer.

Zhuge, H., & Zhang, J. (2011). Automatically constructing semantic link network on documents. *Concurrency and Computation*, *23*, 956–971. doi:10.1002/cpe.1624

KEY TERMS AND DEFINITIONS

Concept: One or multiple words associated with a category that was generated by the abstraction of common characteristics from a range of particular ideas, while removing the uncommon characteristics. The remaining common characteristic is that which is similar to all of the different individuals and represents the meanings, or sense, of the ideas.

Concept-Based Indexing (CBI): A method for indexing that differs from text-based indexing (which uses keywords or headings); CBI instead uses descriptions, ideas, and concepts to index documents.

Semantic Core: The document-specific component of the semantic network that contains the ideas, concepts, that best represents the meaning of the document, rather than the best-matching words.

Concept Retrieval: The ability to query a document and extract particular segments of text that match concepts or ideas provided by a user.

Semantic Analysis: The elicitation of knowledge from documents, accounting for the context and understanding. The units that are extracted are arranged and grouped within meaningful categories.

Semantic Document Network: A network that contains the semantic representation of content of the document but not the document textual content. It is the intersection between the content of the documents and connects the nodes, representing the overlap of semantic content of documents.

Semantic Network: Nodes, encapsulating data and information, are connected by edges which include information about how these nodes are related to one another.

Text Analysis: The process of deriving meaningful information from the data and ideas expressed within the document. It includes meta-information, structural information, and content information.