Title	Representation of Ontology-Based Biological Information Extraction
All Authors	Khin Myo Sett1 , and Nwe Nwe Win2
Publication Type	International publication
Publisher (Journal name, issue no., page no etc.)	Journal of the Myanmar Academy of arts and science, Vol. XIII, No. 3
Abstract	An information extraction (IE) system has been developed by using ontology for extracting biological information. Ontologies in IE may provide new techniques for supporting open tasks of semantic analyses regarding for instance temporal analyses, resolution of contradiction, or context awareness. The system uses knowledge representation techniques for extracting information. Graph-based representation of data can result in effective and scalable methods for information extraction.
Keywords	Information extraction, Ontology, Graph Database
Citation	
Issue Date	2015

Representation of Ontology-Based Biological Information Extraction

Khin Myo Sett[1] , and Nwe Nwe Win[2]

Abstract

An information extraction (IE) system has been developed by using ontology for extracting biological information. Ontologies in IE may provide new techniques for supporting open tasks of semantic analyses regarding for instance temporal analyses, resolution of contradiction, or context awareness. The system uses knowledge representation techniques for extracting information. Graph-based representation of data can result in effective and scalable methods for information extraction.

Key words: Information extraction, Ontology, Graph Database

1. Introduction

Ontologies are formal, explicit specifications of a shared conceptualization. This means that ontologies are useful to model knowledge in a formal abstract way which can be read by computers. With ontologies it is possible to represent concepts, relations among concepts and even constraints on their use. Annotations are a linkage between the knowledge and contents. On one hand, knowledge is represented by means of ontologies. On the other hand, contents are pieces of raw text that need a meaning and which are linked with ontological concepts. Due to the interest in automated analysis of all this information, in recent years, there has been a growing interest in the research community in developing data mining techniques, such as knowledge-based data mining and classification which are able to exploit this kind of information. They are typically applied over structured textual attributes which correspond to features of the analyzed entities. In these cases, attribute labels (i.e., words or noun phrases) are interpreted by mapping them to concepts and analyzing the background knowledge structure to which these concepts belong. However, these methods are rarely able to deal with raw text, from which relevant

[1] 4PhD-Re-Com-1, Lecturer, Department of Computer Studies, Dagon University

[2] Professor and Head, Retired, Dr., Department of Computer Studies, Yangon University

features should be extracted and matched to ontological entities before the data analysis.

Ontology defines the common terms and concepts (meaning) used to describe and represent an area of knowledge. An ontology can range in expressivity from a taxonomy (knowledge with minimal hierarchy or a parent/child structure), to a thesaurus (words and synonyms), to a conceptual model (with more complex knowledge), to a logical theory (with very rich, complex, consistent and meaningful knowledge).

In computer science, ontologies are graphs in which their nodes (i.e. vertices) are entities or concepts, and edges correspond to semantic relations between them. Ontology definition languages offer a wide set of constructors allowing complex ontology modeling. The most common type of relations used in ontologies are hypernymy (i.e. is-a) and meronymy (i.e. is-part-of or its inverse has-part). The former allows to define an entity class hierarchy and the latter describes the properties of each concept. Other important quality of the ontologies is its ability to obtain new knowledge using logical inference or reasoning. The knowledge included in domain ontologies describes objects and domains with concepts and relations.

2. Information Extraction

There has been an explosive growth in the amount of information available on networked computers around the world, much of it in the form of natural language documents. Information Extraction (IE) is the task of locating specific pieces of data within a natural language document. Moreover, the advent of the internet has given IE a particular commercial relevance. IE is a process which takes unseen texts as input and produces fixed format, unambiguous data as output. At the core of an IE system is an extractor, which processes text; it overlooks irrelevant words and phrases and attempts to home in on entities and the relationships between them. These data may be used directly for display to users, or may be stored in a database or spread sheet for direct integration with a back-office system, or may be used for indexing purposes in search engine/Information Retrieval (IR) applications. The goal of Information Extraction (IE) is to automatically extract structured information from unstructured text sources. An IE system has to solve two subtasks: (1) Entity Extraction: identifying the strings *”Lion”* and *”Animal”* as entities and assigning semantic classes to them (Lion and Animal, respectively); (2) Relation Extraction: identifying the binary is-a relation and its two arguments. The task of assigning semantic classes to instances (e.g., assigning class mammal to*”Lion”*) can be thought of as extraction of another special binary relation: the IS-A relation.

3. Ontology and Information Extraction

In recent years, ontologies have emerged as a new paradigm to model and formalize domain knowledge in a machine readable way. An ontology is defined as “a formal, explicit specification of a shared conceptualization”. Conceptualization refers to an abstract model of some phenomenon in the world by having identified its relevant concepts.

Ontologies are designed for being used in applications that need to process the content of information, as well as to reason about it, instead of just presenting information to humans. They permit greater machine interpretability of content than that supported by XML, RDF and RDF Schema (RDF-S), by providing additional vocabulary along with a formal semantics. So, ontologies represent an ideal knowledge background in which to base text understanding and enable the extraction of relevant information. This may enable the development of more flexible and adaptive IE systems than those relying on manually composed extraction rules (both based on linguistic constructions or document structure).

Oontologies can assist both manually or semi-automatically constructed rule-based IE systems. On the one hand, the knowledge engineer can commit to the ontology, which would guarantee that the extraction rules are tailored to extract the kind of information represented in the ontology. On the other hand, an annotator can commit to the ontology and annotate only parts of text that are relevant from the ontology’s point of view.

4. Ontology exploitation for IE

IE and ontologies are involved in two main and related tasks :

- Ontology is used for Information Extraction: IE needs ontologies as part of the understanding process for extracting the relevant information;
- Information Extraction is used for populating and enhancing the ontology: texts are useful sources of knowledge to design and enrich ontologies.

These two tasks, as can be seen in Figure 1, can be combined in a cyclic process: ontologies are used for interpreting the text at the right level for IE and IE extracts new knowledge from text, to be integrated in the ontology.

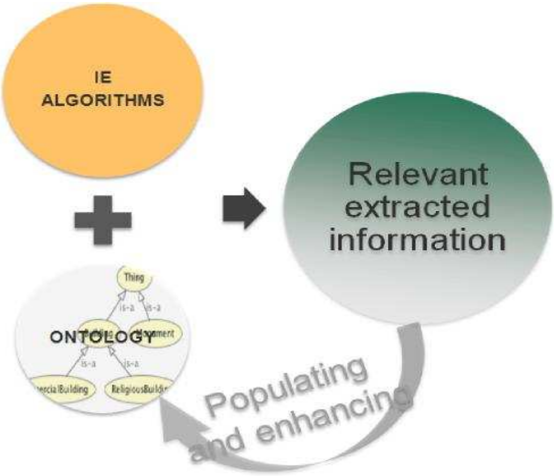


Figure 1. Ontology exploitation for IE (cyclic process)

5. Graph Database and Bioinformatics

Graph database have seen extensive application in the field of bioinformatics. Bioinformatics use graph databases to relate a complex web of information that genes, proteins and enzymes, etc. It is open source and the intention is to be easy to expand upon. Gathering huge amounts of complex information (data and knowledge) is very common nowadays. It is necessity to represent, store and manipulate complex information (e.g. detect correlations and patterns, discover explanations, construct predictive models etc.). Furthermore, being autonomously maintained, data can change in time or even change its base structure, making it difficult for representation systems to accommodate these changes. Current representation and storage systems are not very flexible in dealing with big changes and also they are not concerned with the ability of performing complex data manipulations of the sort mentioned above. On the other hand, data manipulation systems cannot easily work with structural or relational data, but just with flat data representations. To bridge the gap between these two, by introducing a new type of database structure, called Graph Databases (GDB), based on a natural graph representation. Our Graph Databases are able to represent as graphs *any kind of information*, naturally accommodate changes in data, and they also make easier for Machine Learning methods to use the stored information.

A graph is a data structure composed of edges and vertices. Graph database technology is an effective tool for modeling data when a focus on the relationship between entities is a driving force in the design of a data model. Modeling objects and the relationships between them means almost anything can be represented in a corresponding graph. A common graph type supported by most systems is the property graph. Property graphs are attributed, labeled, directed multi-graphs. Figure 2 provides a visual example of a property graph which represents interactions between people and objects. A benefit to the multi graph is that it is the most complex implementation because every other type of graph consists of subsets of the property graph implementation. This means a property graph can effectively model all other graph types. The graph database is optimized for the efficient processing of dense, interrelated datasets. This design allows the construction of predictive models, and detection of correlations and patterns. This highly dynamic data model in which all nodes are connected by relations allows for fast traversals along the edges between vertices. A particular benefit is the fact that traversals are localized and do not have to take into account sets of unrelated data. A problem that is inherent in SQL.

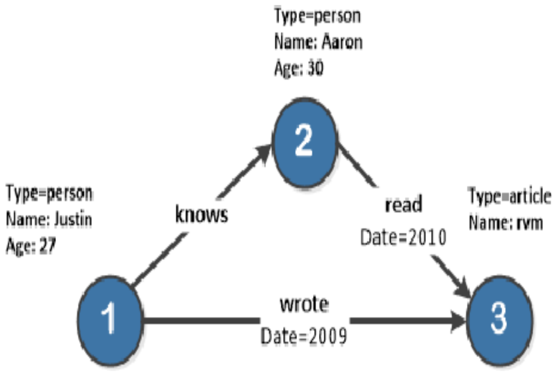


Fig 2. Example of property Graph

5.1 Core Concepts

A graph is an object which contains nodes and relationships. Nodes have properties and are organized by relationships which also have properties. A traversal navigates a graph and identifies paths which order nodes. Figure 3 illustrates the individual components of a graph and gives a visual representation of how they relate to one another.

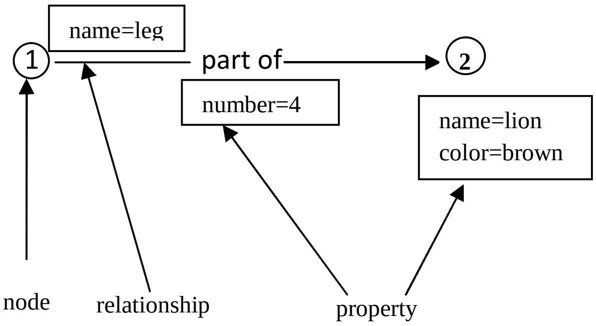
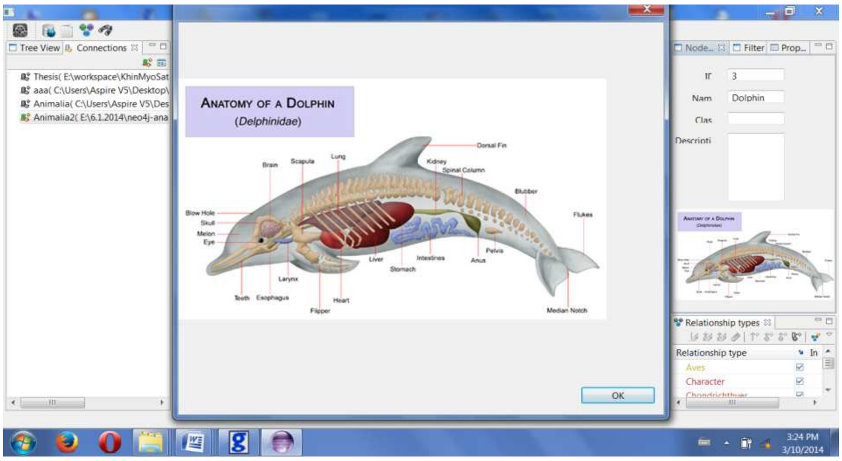


Fig .3. Individual Component of Graph

6. Ontology-Based Information Extraction

The ontology-based information extraction uses knowledge representations which combine the features of text based retrieval and content-based retrieval. The main purpose on ontology is to represent the information in semantic manner. Hence, the information is represented in machine understandable manner, this leads to extract task easier. We have created a general ontology hierarchy by using Graph Database, because of the structures of ontology and graph database are the same. And they are hierarchical and tree-like structures. For example, the collection of image is having about 2000 images. We have categorized the things into two major divisions. The images are mapped to ontology with complete relationship and content description of the image. For example, since “mammals” is a subcategory of “animals”, an image annotated with the keyword “mammals” is found using the keyword “animals”. The same idea of enlarging a keyword with related terms in order enhances access rate.


Thesis(E:\workspace\KhinMyoSat
aaa(C:\Users\Aspire V5\Desktop\
Animalia(C:\Users\Aspire V5\Des
Animalia2(E:\6.1.2014\neo4j-ana

Ir 3
Nam Dolphin
Clas
Descripti

Relationship type	In
Aves	☑
Character	☑
Chondrichthyes	☑

Fig.4. Extracted Image

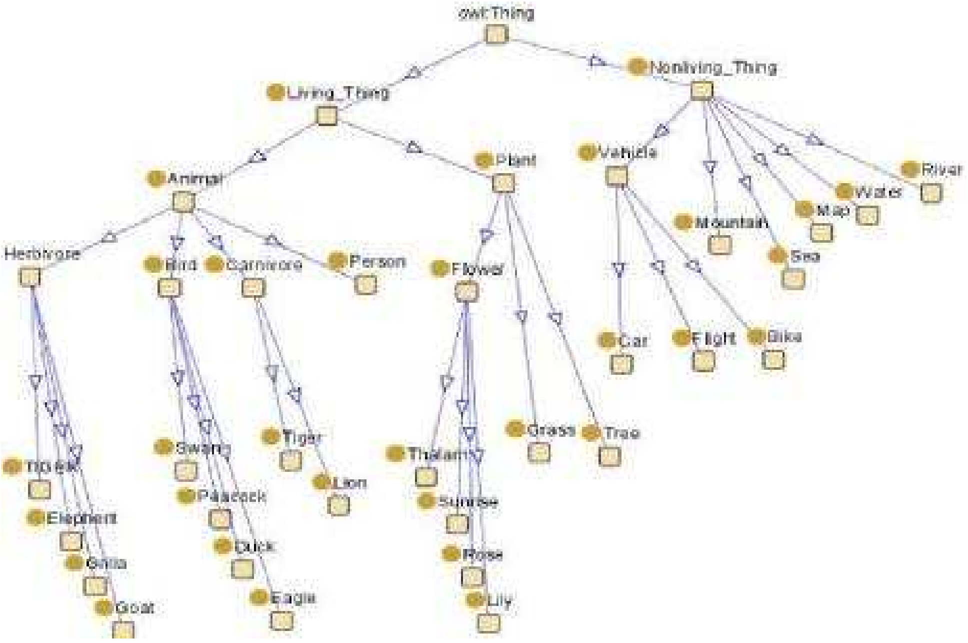


Fig. 5. Ontology Structure

The image then inherits the class properties and annotation of its assigned class. It is based on the idea of image annotation using ontologies. High-level concepts are efficiently stored and automatically mapped to visual features or objects which are extracted by various image or information analysis techniques. The ICONCLASS is used in extracting the low level visual features of an image.

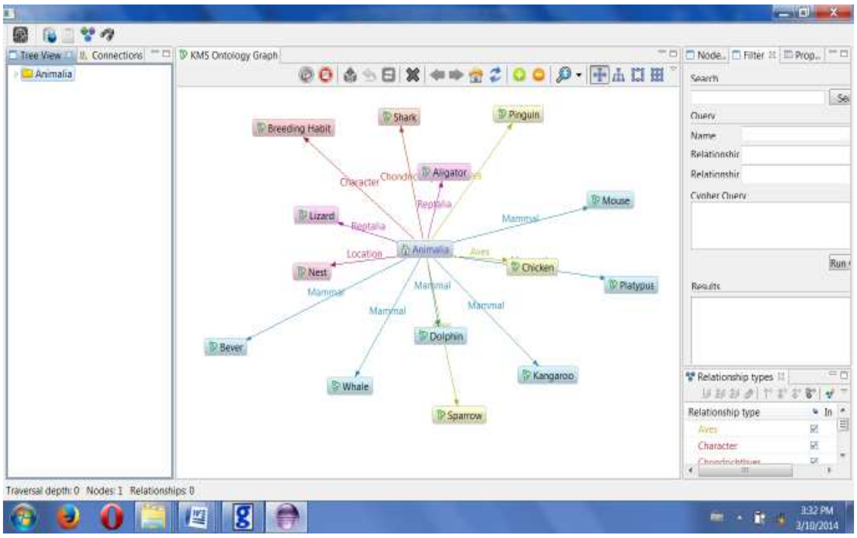


Fig .6. Visualization of Graph Database

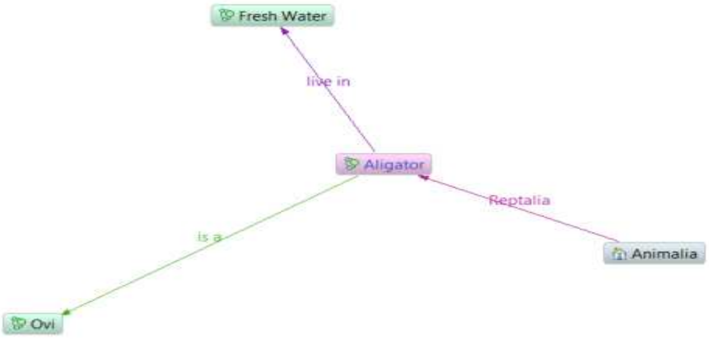
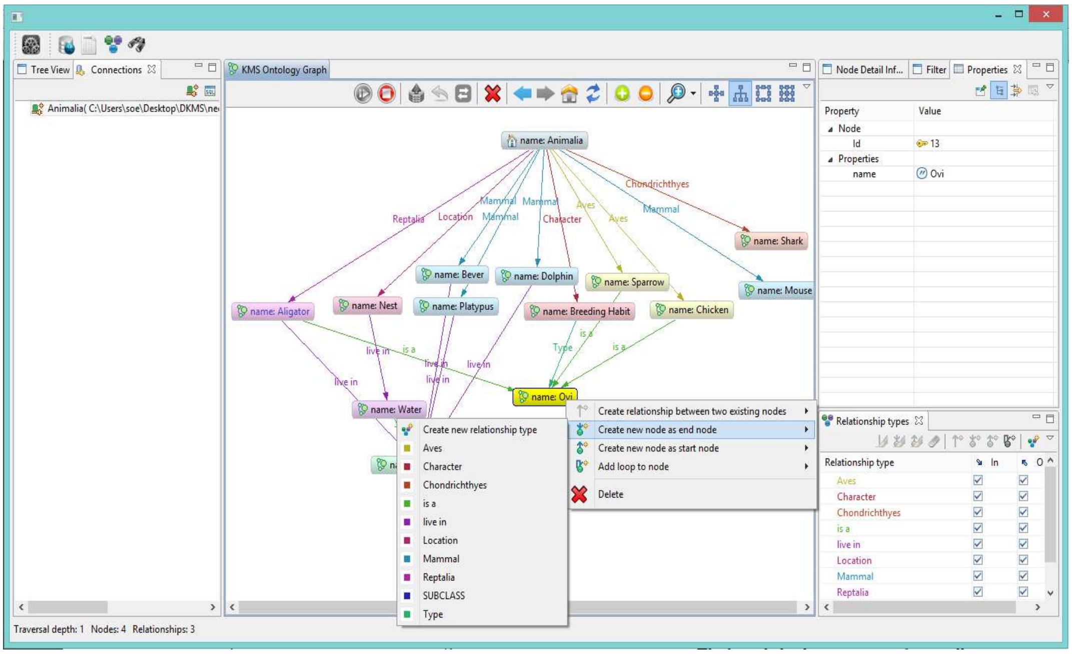


Fig.7. Conceptual Data Model

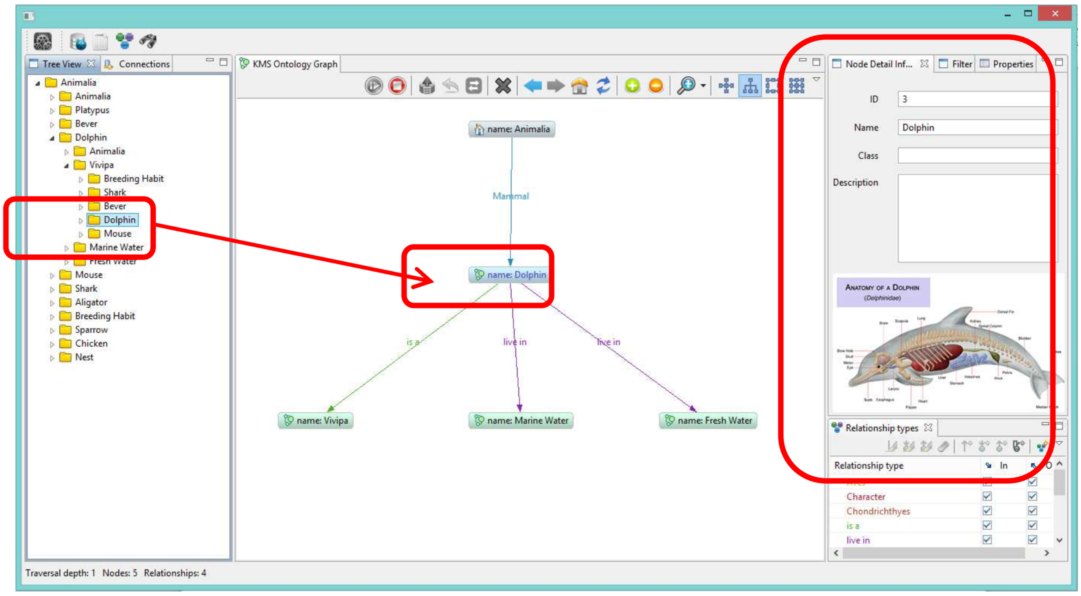


Property	Value
Node	
Id	13
Properties	
name	Ovi

Relationship type	In	O
Aves	☑	☑
Character	☑	☑
Chondrichthyes	☑	☑
is a	☑	☑
live in	☑	☑
Location	☑	☑
Mammal	☑	☑
Reptalia	☑	☑

Traversal depth: 1 Nodes: 4 Relationships: 3

Fig.8. Creating New Node



ID 3

Name Dolphin

Class

Description

Relationship type	In	O
Aves	☑	☑
Character	☑	☑
Chondrichthyes	☑	☑
is a	☑	☑
live in	☑	☑

Traversal depth: 1 Nodes: 5 Relationships: 4

Fig. 9. Browsing Items via Tree Structure and Graph Structure

7. Conclusion

Introducing ontology in the collection of information e.g images with complete relationship enhances the retrieval of images or information at a faster rate when compared to ordinary retrieval. Graph database provides the retrieval of images or information. Accuracy has been achieved through the ontological structure. Ontology-enriched knowledge base of metadata can be applied to constructing more meaningful answers to queries.

Acknowledgement

I would like to express my sincere grateful to Professor Dr. Nwe Nwe Win, Head of Department of Computer Studies (Retired), University of Yangon for effective guidance on my research. I would like to thank Dr. Soe Mya Mya Aye, Associate Professor, Head of Department of Computer Studies, University of Yangon for her kind permission to carry out this research.

References

1. Davies J., Studer R., Warren P.. Semantic Web Technologies Trends and Research in Ontology-based Systems. John Wiley & Sons Ltd, 2006.

2. Eugenio Di Sciascio, Francesco M.Donini, Marina. Structured Knowledge Representation for Image Retrieval BARI Italy.

3. Hull, R. and King, R. 1987. Semantic database modeling: Survey, applications, and research issues. *ACM Comput. Surv. 19,* 3, 201-260. number 6, pages 35-41, doi:10.1002/bult.2010.1720360610, ISSN:1550 held in 19-20 March 2009 at IRTT, Erode. Department of Computer Science, P.O. Box.

4. Miller. Justin J. Graph Database Applications and Concepts with Neo4j, Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA March 23rd-24th, 2013.

5. Miller. Justin J. Mukerji (eds.)**.** MDA Guide Version 1.0. OMG Document: omg/2003-05-01**.** http://www.omg. org/docs/omg/03-05-01.pdf (2007-03-20).

6. N.Magesh. Knowledge Based Approach for Language translation Organized by Computer Science and Engineering at Vellalar College of Engg and Technology Erode02. National Conference on Recent Trents in Innovative Technologies on Novenber 2009.

7. P. Mitra, G. Wiederhold, M. Kersten**.** A graph oriented model for articulation of ontology interde-pendencies. Proc. Extending DataBase Technologies, Springer, Berlin Heidelberg, 2000, LNCS 1777, 86–100.

8. Rodriguez, M.A., Neubauer, P. Constructions from Dots and Lines Bulletin of the American Society for Information Science and Technology, American Society for Information Science and Technology, volume 36, 8366, August 2010.

9. Silvescu, Adrian and Caragea, Doina and Altramentov, Anna *Graph Databases.* Iowa State University.

10. Van Rijsbergen, Information Retrieval, London: Butterworths, Second Edition.