Structural and Content Queries on the Nested Sets Model

Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
silvello@dei.unipd.it

Abstract. Hierarchical structures are pervasive in computer science. They are a fundamental means for modeling many aspects of reality and for managing wide corpora of data and digital resources. How to model and manage data hierarchies is a major theme in database theory and practice. The most important hierarchical structure is the tree, which has been widely studied, analyzed, and adopted in several contexts and scientific fields over time. In this paper we describe an alternative approach for modeling and managing hierarchies, called the Nested Sets Model (NS-M) and we describe how it is possible to reconsider structural and content queries by exploiting this model.

1 Motivation

Hierarchical structures are pervasive in computer science. They are a fundamental means for modeling many aspects of reality and for managing wide corpora of data and digital resources. How to model and manage data hierarchies is a major theme in database theory as well as in database applications [1] and it has been extensively treated in the context of database abstractions, the relational model, the complex value model, object-oriented database systems, and semistructured data and Extensible Markup Language (XML).

In the database field and more generally in computer science, the fundamental structure used to model and represent hierarchies is the tree. The tree has been formalized and deeply studied, its properties are well understood, and many of the algorithms based on it run in polynomial time. Therefore, the tree has been exploited by researchers and developers in many different fields to model and solve their problems. In this sense, trees are considered “the most important nonlinear structures that arise in computer science” [6].

In this article we propose an alternative approach for modeling and managing hierarchies, called Nested Sets Model (NS-M), which is based on the inclusion between sets as a means to capture hierarchical relations. The foundational idea underlying this model is to express the hierarchical relationships between objects through the inclusion property between sets, in place of the binary relation between nodes exploited by the tree. Therefore, the pivotal question we want to address is how structural and content queries on hierarchies can be handled with a data model other than the tree. In order to answer this question we define the
Preface

This volume collects the papers and posters selected for presentation at the Twentieth Italian Symposium on Advanced Database Systems (SEBD 2012), held in Venice from the 24th to the 27th of June 2012.

The first edition of SEBD was held in 1993 as an outcome of a nationwide project funded by the Italian National Research Council, called Logidata+. The success of the event encouraged the organizers to repeat the experience year after year, until the symposium has became the annual gathering where the Italian database research community discusses not only original research, but also trends, policies, funding, assessment, and other cogent problems of research and teaching in computer science in general.

This twentieth birthday of SEBD represents a major achievement of the Italian database research community, which has shown its ability to react to, and cope with, the many research challenges that have been arising during these twenty years. Together, we have witnessed the activities and liveliness of a research community that has involved “generations” of researchers, offering to young researchers, through SEBD, a stage for discussing and confronting ideas.

Traditionally, the database research community has mainly concentrated on the methodologies, techniques, and technologies for data management. While in the past the role of data was recognized as central mostly in the organization of business activities, in the last decade this equilibrium has been completely overturned, and data-related research and development problems are involved in every individual and collective activity of society.

The main two phenomena that have nurtured this trend are the availability of information through the Web and the possibility to access resources anywhere and at any time through mobile, portable devices. Now everybody can both consume and produce information by means of various sorts of easy interaction means: social computing involves people and stimulates them to the use of sophisticated applications and devices; the pervasiveness of computation means (very small, cheap, and low power devices, connected through wireless networks) relieves people of a cumbersome interaction with passive tools, by embedding the processing power in the environment, often anticipating the user’s needs; new forms of information, like scientific or multimedia data, require equally new forms of storage and processing. These are only examples of the most various information, from multimedia data streams and storage systems to semantic-web knowledge linked all over the Web, to natural language information that has to be automatically understood.

The availability of so huge amounts of very different kinds of data (“big data”) is changing our attitude towards them, requiring more and more sophisticated analysis means. Therefore, a data-centric vision of the world is actually key for advancing in such a demanding scenario. The vision of an inclusive society for the future requires effective, usable, and secure methodologies and techniques
for data production, management and analysis, which will enable us to give semantics to, and make sense of, the enormous quantity of information which is continuously produced and very often ignored or ill-used.

In this scenario, two apparently contrasting phenomena take place: on the one hand, the research becomes more and more specialized as the results of investigation on long-term research problems generate the need to delve deeper into more specific issues; on the other hand, different disciplines join forces with ours, displaying their full power to answer demands posed by brand-new applications and visionary technologies.

These two trends are shown by the topics presented in this edition of SEBD: Knowledge Discovery and Extraction; Data Streams and Uncertainty; Semantic Web and Interoperability; Data Mining; Information Access and Retrieval; Query Answering; Logic and Deductive Databases; Classification and Clustering; and, Integration and Information Exchange.

37 contributions were originally submitted, of which 9 full papers and 28 extended abstracts. Most of the contributions came from universities and the CNR (Italian National Research Council). Each submission was evaluated by three independent referees and the selection process led to the acceptance of 4 full papers, 23 extended abstracts, and 7 posters. Posters, which were introduced for the first time in the 2011 edition of SEBD, represent a better mean for either discussing new emerging ideas or presenting more consolidated research in a more interactive and less time-constrained fashion.

Besides paper and poster presentations, the programme also featured two keynote talks which highlighted future research directions. The first one by prof. Timos Sellis (Research Center “Athena” and National Technical University of Athens, Greece) explored “Personalization in Web Search and Data Management”; the second one by Dr. Krishna P. Gummadi (Max Planck Institute for Software Systems (MPI-SWS), Germany) dealt with “Extracting Relevant and Trustworthy Information from Microblogs”. Finally, one tutorial by prof. Maarten de Rijke (ISLA, University of Amsterdam, The Netherlands) concerned “Log File Analysis and Mining”. Altogether, the invited talks and the tutorial not only offered an interesting perspective on compelling research challenges but also provided bridges connecting typical database themes with topics of nearby disciplines, such as information access and retrieval.

Moreover, consistently with the already-mentioned custom of discussing important issues which are transversal with respect to the technical topics, the symposium hosted two “Community Think-Tank” sessions, during which the fundamental topic of research evaluation was discussed.

We wish to thank all the authors who submitted papers and all the conference participants. We are also grateful to the members of the Programme Committee and the external referees for their thorough work in reviewing submitted contributions with expertise and patience, and to the members of the SEBD Steering Committee for their support in the various decisions that had to be taken during the event organization. Finally, we wish to thank the members of the Local
Organization Committee for their willingness and contribution in dealing with the many aspects which are involved by the logistics of a conference.

We gratefully acknowledge the patronage of several institutions: Regione del Veneto (http://www.regione.veneto.it/); Università degli Studi di Padova (http://www.unipd.it/); and, Dipartimento di Ingegneria dell’Informazione dell’Università degli Studi di Padova (http://www.dei.unipd.it/).

We also express our gratitude to the many sponsoring organizations and initiatives for their significant and timely support: Associazione Italiana per l’Informatica ed il Calcolo Automatico (AICA, http://www.aicanet.it/); Dipartimento di Ingegneria dell’Informazione dell’Università degli Studi di Padova; PROMISE (FP7 NoE, contract n. 258191, http://www.promise-noe.eu/); and, CULTURA (FP7 STREP, contract n. 269973, http://www.cultura-strep.eu/).

This SEBD edition comes only a few weeks after the earthquake that hit the Italian region of Emilia Romagna in May 2012. The SEBD community wishes to express its empathy with the many region inhabitants who have lost their homes, jobs and even lives, making good omens for a quick reconstruction of the social, productive, and cultural life in that area.

June 2012
Nicola Ferro
Letizia Tanca
Organization

SEBD 2012 is organized by the Information Management Systems (IMS) research group of the Department of Information Engineering (DEI) of the University of Padua, Italy.

General Chair

Letizia Tanca, Politecnico di Milano

Program Chair

Nicola Ferro, Università degli Studi di Padova

Local Organization Chair

Gianmario Silvello, Università degli Studi di Padova

Local Organization Committee

Debora Leoncini, Università degli Studi di Padova
Ivano Masiero, Università degli Studi di Padova
Simone Peruzzo, Università degli Studi di Padova

Program Committee

Giuseppe Amato, Istituto di Scienza e Tecnologie dell’Informazione, CNR Pisa
Elena Baralis, Politecnico di Torino
Carlo Batini, Università degli Studi di Milano-Bicocca
Devis Bianchini, Università degli Studi di Brescia
Andrea Calì, University of London, Birkbeck College
Cinzia Cappiello, Politecnico di Milano
Michelangelo Ceci, Università degli Studi di Bari
Augusto Celentano, Università Ca’ Foscari Venezia
Paolo Ciaccia, Università di Bologna
Valter Crescenzi, Università degli Studi Roma Tre
Paolino Di Felice, Università degli Studi dell’Aquila
X Organization

Claudia Diamantini, Università Politecnica delle Marche
Alfio Ferrara, Università degli Studi di Milano
Enrico Franconi, Libera Università di Bolzano
Giorgio Ghelli, Università di Pisa
Giovanna Guerrini, Università degli Studi di Genova
Nicola Orio, Università degli Studi di Padova
Luigi Palopoli, Università della Calabria
Antonella Poggi, Università degli Studi di Roma “La Sapienza”
Carlo Santiani, Università degli Studi della Basilicata
Domenico Ursino, Università Mediterranea di Reggio Calabria
Yannis Velegrakis, Università degli Studi di Trento
Maurizio Vincini, Università degli Studi di Modena e Reggio Emilia

External Reviewers

Domenico Beneventano, Università degli Studi di Modena e Reggio Emilia
Mirko Bronzi, Università degli Studi Roma Tre
Luca Caglierio, Politecnico di Torino
Dario Colazzo, Université Paris Sud
Fabio Fassetti, Università della Calabria
Alessandro Fiori, Politecnico di Torino
Sergio Flesca, Università della Calabria
Giancarlo Fortino, Università della Calabria
Paolo Garza, Politecnico di Milano
Lorenzo Genta, Università degli Studi di Milano
Alberto Grand, Politecnico di Torino
Gianluca Lax, Università Mediterranea di Reggio Calabria
Lucrezia Macchia, Università degli Studi di Bari
Giuseppe Manco, Istituto di Calcolo e Reti ad Alte Prestazioni, CNR Cosenza
Andrea Marin, Università Ca’ Foscari Venezia
Andrea Maurino, Università degli Studi di Milano-Bicocca
Marek Maurizio, Università Ca’ Foscari Venezia
Michele Melchiori, Università degli Studi di Brescia
Paolo Merialdo, Università degli Studi Roma Tre
Marco Mesiti, Università degli Studi di Milano
Luigi Moccia, Istituto di Calcolo e Reti ad Alte Prestazioni, CNR Cosenza
Antonino Nocera, Università Mediterranea di Reggio Calabria
Marius-Octavian Olaru, Università degli Studi di Modena e Reggio Emilia
Matteo Palmonari, Università degli Studi di Milano-Bicocca
Gianvito Pio, Università degli Studi di Bari
Giovanni Quattrone, University College London
Domenico Rosaci, Università Mediterranea di Reggio Calabria
Donatello Santoro, Università degli Studi della Basilicata
Daniela Stojanova, Jozef Stefan Institute Ljubljana

SEBD Steering Committee

Maristella Agosti, Università degli Studi di Padova
Antonio Albano, Università di Pisa
Paolo Atzeni, Università degli Studi Roma Tre
Sonia Bergamaschi, Università degli Studi di Modena e Reggio Emilia
Valeria De Antonellis, Università degli Studi di Brescia
Maurizio Lenzerini, Università degli Studi di Roma “La Sapienza”
Domenico Saccà, Università della Calabria
Fabio Schreiber, Politecnico di Milano
Letizia Tanca, Politecnico di Milano
Paolo Tiberio, Università degli Studi di Modena e Reggio Emilia
# Table of Contents

## Keynote Addresses

- Extracting Relevant and Trustworthy Information from Microblogs  
  *Krishna P. Gummadi*  
  Page 1
- Personalization in Web Search and Data Management  
  *Timos Sellis (joint work with T. Dalamagas, G. Giannopoulos, and A. Arvanitis)*  
  Page 3

## Tutorials

- Log File Analysis and Mining  
  *Maarten de Rijke*  
  Page 5

## Community Think Tank

- Sulla Classificazione delle Sedi di Pubblicazione nella Valutazione della Produzione Scientifica  
  *Giansalvatore Mecca, Marcello Buoncristiano, and Donatello Santoro*  
  Page 7

## Knowledge Discovery and Extraction

- Discovering Hidden *me* Edges in a Social Internetworking Scenario  
  *Francesco Buccafurri, Gianluca Lax, Antonino Nocera, and Domenico Ursino*  
  Page 15
- Count Constraints for Inverse OLAP and Aggregate Data Exchange  
  *Domenico Saccà, Edoardo Serra, and Antonella Guzzo*  
  Page 27
- Individual Mobility Profiles: Methods and Application on Vehicle Sharing  
  *Roberto Trasarti, Fabio Pinelli, Mirco Nanni, and Fosca Giannotti*  
  Page 35

## Data Streams and Uncertainty

- Getting the Best from Uncertain Data: the Correlated Case  
  *Ilaria Bartolini, Paolo Ciaccia, and Marco Patella*  
  Page 43
- Relaxed Queries over Data Streams  
  *Barbara Catania, Giovanna Guerrini, Maria Teresa Pinto, and Paola Podestà*  
  Page 51
- A Logic-based Language for Data Streams  
  *Carlo Zaniolo*  
  Page 59
## Table of Contents

### Semantic Web and Interoperability

A Framework for Biological Data Normalization, Interoperability, and Mining for Cancer Microenvironment Analysis ........................................... 67  
*Michelangelo Ceci, Mauro Coluccia, Fabio Fumarola, Pietro Hiram Guzzi, Federica Mandreoli, Riccardo Martoglia, Elio Masciari, Massimo Mecella, and Wilma Penzo*

A Lightweight Model for Publishing and Sharing Linked Web APIs ....... 75  
*Devis Bianchini, Valeria De Antonellis, and Michele Melchiori*

A Framework for Building Multimedia Ontologies from Web Information Sources ................................................................. 83  
*Angelo Chianese, Vincenzo Moscato, Fabio Persia, Antonio Picariello, and Carlo Sansone*

### Integration and Information Exchange

Integration and Provenance of Cereals Genotypic and Phenotypic Data 91  
*Domenico Beneventano, Sonia Bergamaschi, and Abdul Rahman Dannaoui*

A Short History of Schema Mapping Systems ................................. 99  
*Giansalvatore Mecca, Paolo Papotti, and Donatello Santoro*

A Linear Algebra Approach for Supply Chain Management ............... 107  
*Ricoerto De Virgilio and Franco Milicchio*

### Information Access and Retrieval

A Semantic Method for Searching Knowledge in a Software Development Context ................................................................. 115  
*Sonia Bergamaschi, Riccardo Martoglia, and Serena Sorrentino*

An Efficient Methodology for the Identification of Multiple Music Works within a Single Query ......................................................... 123  
*Emanuele Di Buccio, Nicola Montecchio, and Nicola Orio*

Using Keywords to Find the Right Path through Relational Data ....... 131  
*Ricoerto De Virgilio, Antonio Maccioni, and Riccardo Torlone*

### Query Answering

Efficient Diversification of Top-k Queries over Bounded Regions ....... 139  
*Piero Fraternali, Davide Martinenghi, and Marco Tagliasacchi*

A Query Reformulation Framework for P2P OLAP .......................... 147  
*Matteo Golfarelli, Federica Mandreoli, Wilma Penzo, Stefano Rizzi, and Elisa Turricchia*
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficient Query Answering over <em>Datalog</em> with Existential Quantifiers</td>
<td>155</td>
</tr>
<tr>
<td>Nicola Leone, Marco Manna, Giorgio Terracina, and Pierfrancesco Veltri</td>
<td></td>
</tr>
<tr>
<td>Logic and Deductive Databases</td>
<td></td>
</tr>
<tr>
<td>The Algebra and the Logic for SQL Nulls</td>
<td>163</td>
</tr>
<tr>
<td>Enrico Franconi and Sergio Tessaris</td>
<td></td>
</tr>
<tr>
<td>An Extension of <em>Datalog</em> for Graph Queries</td>
<td>177</td>
</tr>
<tr>
<td>Mirjana Mazuran, Edoardo Serra, and Carlo Zaniolo</td>
<td></td>
</tr>
<tr>
<td>Stratification-based Criteria for Checking Chase Termination</td>
<td>185</td>
</tr>
<tr>
<td>Sergio Greco, Francesca Spezzano, and Irina Trubitsyna</td>
<td></td>
</tr>
<tr>
<td>Classification and Clustering</td>
<td></td>
</tr>
<tr>
<td>Approximation of the Gradient of the Error Probability for Vector</td>
<td>193</td>
</tr>
<tr>
<td>Quantizers</td>
<td></td>
</tr>
<tr>
<td>Claudia Diamantini, Laura Genga, and Domenico Potena</td>
<td></td>
</tr>
<tr>
<td>Topic Modeling for Segment-based Documents</td>
<td>205</td>
</tr>
<tr>
<td>Giovanni Ponti, Andrea Tagarelli, and George Karypis</td>
<td></td>
</tr>
<tr>
<td>Knowledge Discovery from Textual Sources by using Semantic Similarity</td>
<td>213</td>
</tr>
<tr>
<td>Paolo Atzeni, Fabio Polticelli, and Daniele Toti</td>
<td></td>
</tr>
<tr>
<td>Data Mining</td>
<td></td>
</tr>
<tr>
<td>Mining Temporal Evolution of Criminal Behaviors</td>
<td>221</td>
</tr>
<tr>
<td>Gianvito Pio, Michelangelo Ceci, and Donato Malerba</td>
<td></td>
</tr>
<tr>
<td>Privacy-preserving Mining of Association Rules from Outsourced</td>
<td>233</td>
</tr>
<tr>
<td>Transaction Databases</td>
<td></td>
</tr>
<tr>
<td>Fosca Giannotti, Laks V.S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui Wang</td>
<td></td>
</tr>
<tr>
<td>Frequent Itemset Mining of Distributed Uncertain Data under User-Defined Constraints</td>
<td>243</td>
</tr>
<tr>
<td>Alfredo Cuzzocrea and Carson K. Leung</td>
<td></td>
</tr>
<tr>
<td>Posters</td>
<td></td>
</tr>
<tr>
<td>Hierarchical Latent Factors for Preference Data</td>
<td>251</td>
</tr>
<tr>
<td>Nicola Barbieri, Giuseppe Manco, and Ettore Ritacco</td>
<td></td>
</tr>
<tr>
<td>Integrating Clustering Techniques and OLAP Methodologies: The ClustCube Approach</td>
<td>257</td>
</tr>
<tr>
<td>Alfredo Cuzzocrea and Paolo Serafino</td>
<td></td>
</tr>
</tbody>
</table>
Enhancing Datalog with Epistemic Operators to Reason About Knowledge in Distributed Systems ............................................... 265  
  Matteo Interlandi

On Casanova and Databases or the Similarity Between Games and DBs.. 271  
  Giuseppe Maggiore, Renzo Orsini, and Michele Bugliesi

On the Use of Dimension Properties in Heterogeneous Data Warehouse Integration ............................................................ 277  
  Marius-Octavian Olaru and Maurizio Vincini

Structural and Content Queries on the Nested Sets Model....................... 283  
  Gianmaria Silvello

Knowledge-based Real-Time Car Monitoring and Driving Assistance .... 289  
  Michele Ruta, Floriano Scioscia, Filippo Gramegna, Giuseppe Loseto, and Eugenio Di Sciascio

Author Index .............................................................. 295
NS-M (Section 2.1), we describe how it allows us to handle structural and content components of a hierarchy (Section 2.2) and we prove its properties (Section 2.3). In Section 3 we specify basic structural and content query operations by highlighting how they can be handled by the NS-M; and, in Section 4 we draw some final remarks and future works.

2 The Nested Sets Model (NS-M)

2.1 Definition of the Model

The NS-M is formally defined as a collection of subsets where specific conditions must hold [2,4].

Definition 1 Let $A$ be a set and let $C$ be a collection of subsets of $A$. Then $C$ is a Nested Sets Collection (NS-C) if:

\begin{align*}
\forall H, K \in C \mid H \cap K \neq \emptyset \Rightarrow H \subseteq K \lor K \subseteq H. \quad \text{(2.2)}
\end{align*}

Therefore, we define a NS-C as a collection of subsets where two conditions must hold. The first condition (2.1) states that set $A$ which contains all the subsets of the collection must belong to the NS-C itself. The second condition states the intersection of every couple of sets in the NS-C is not the empty-set only if one set is a proper subset of the other one.\(^1\)

The NS-C is represented by means of an Euler-Venn diagram as we can see in Figure 1, which represents a sample NS-C composed of five nested sets:

\(^1\) The graphical representation of a tree as a collection of nested sets was originally presented in [6] with no formal definition of a “nested sets model”; afterwards, this representation has been exploited in [3] to explain an alternative way to solve recursive queries over trees by using an integer intervals encoding. Also [5] proposed the same idea to solve recursion in relational databases.
$C = \{A, B, C, D, E\}$. We can see that $A$ is the top set of $C$ (i.e. the common superset of all the sets in $C$) and thus Condition 2.1 is respected, and all the sets are either disjoints or one is a proper subset of the other as required by Condition 2.2.

2.2 Separating Structure and Content

From the structural point-of-view, a collection of subsets is represented by the sets in the collection and their inclusion dependencies. A collection of subsets at the intensional level is defined by its structure. Let us consider an example, we can say that $C = \{A, B, C\}$ where $B \subseteq A$, $C \subseteq A$ and $B \nsubseteq C \land C \nsubseteq B$ is a NS-C because it respects conditions 2.1 and 2.2 of Definition 1. In this way, we know the structure of the collection and we know which relationships hold between the sets.

When we consider a collection of subsets $C$ from the content point-of-view, it means that we refer to its extensional level. From the content point-of-view a collection of subsets $C$ is represented by the extension of the sets composing it; the properties of the sets are then verified by inspecting the sets and verify the elements that they contain. In this case, we say that the content of a collection of subsets defines the extension of such a collection. Therefore, we can say that $C = \{A, B, C\}$ where $A = \{a, b, c, d\}$, $B = \{b\}$ and $C = \{c, d\}$ is the extension of a NS-C. In the next example we can see a NS-C defined at the intensional level which at the extensional level is satisfied by two different NS-C.

Example 1 Let us consider the following NS-C defined at the intensional level: $C = \{A, B, C, D\}$ where $B \subseteq A$, $C \subseteq A$, $D \subseteq C$ and $B \nsubseteq C \land C \nsubseteq B$. Then, $A = \{a, b, c, d, e\}$, $B = \{b\}$, $C = \{c, d, e\}$, $D = \{d, e\}$ is a valid instance for $C$, as well as $A = \{a, b, c, d, e, f\}$, $B = \{c, d\}$, $C = \{b, e, f\}$ and $D = \{f\}$; indeed, they both satisfy the specified structural conditions.

Both the structural and the content point-of-view are important for the treatment of the NS-M. We exploit the structure defined at the intensional level to define the properties of NS-M; whereas we exploit the extensional level to perform set operations which manipulate the content of the subsets composing the collections.

In the following we make extensive use of the concepts of collection of proper subsets and supersets and of direct subsets and supersets. Let $C$ be a collection of sets and $A \in C$ be a set, we define:

- $S^-(A) = \{B \in C : A \subset B\}$ to be the collection of proper supersets of $A$ in $C$;
- $S^+(A) = \{B \in C : B \subset A\}$ to be the collection of proper subsets of $A$ in $C$.
- $D^-(A) = \{B \in C : (A \subset B) \land (\nexists E \in C \mid A \subset E \subset B)\}$ to be the collection of direct supersets of $A$ in $C$.
- $D^+(A) = \{B \in C : (B \subset A) \land (\nexists E \in C \mid B \subset E \subset A)\}$ to be the collection of direct subsets of $A$ in $C$. 
2.3 Properties of the Model

Many properties of the NS-M are derived by the straightforward application of set theory as we show in the following example which takes into account the intensional level.

**Example 2** Let $C$ be a NS-C. For all $H, K \in C$ \( H \subseteq K \) we can derive from set theory that $H \cup K = K$ and $H \cap K = H$. As well as we can say that for all $H, K \in C$ \( H \nsubseteq K \wedge K \nsubseteq H \Rightarrow H \setminus K = H \wedge K \setminus H = K \).

In this example we see that the sets in a NS-C behave exactly as one would expect under the operations of union, intersection and set difference. Let us see an example which shows how these operations behave at the extensional level.

**Example 3** Let $C = \{A, B, C\}$ be a NS-C, where $B \subseteq A$ and $C \subseteq B$. Then let us consider the following instance: $A = \{a, b, c, d, e\}$, $B = \{c, d, e\}$ and $C = \{e\}$.

Then, $B \cup C = \{c, d, e\} = B$ and $B \cap C = \{e\} = C$.

Let us consider a NS-C $C$, the next proposition shows that for all $H \in C$, $H$ has at most one direct superset.

**Proposition 1** Let $C$ be a NS-C. Then, $\forall H \in C, |D^-(H)| \leq 1$.

**Proof.** Ab absurdo suppose that $\exists H \in C$ such that $|D^-(H)| > 1 \Rightarrow \exists K, L \in D^-(H) \mid H \subseteq K \wedge H \subseteq L \wedge L \nsubseteq K \wedge K \nsubseteq L \Rightarrow H \cap L = H \Rightarrow C$ is not a NS-C (condition 2.2 of Definition 1).

The following corollary to this proposition shows that the set with minimum cardinality in the collection of supersets of $H$ is its direct superset.

**Corollary 2** Let $C$ be a NS-C, $H \in C$ be a set, $S^-(H)$ be the collection of proper supersets of $H$ and $K \in S^-(H)$ where $\forall L \in S^-(H), |K| \leq |L|$ be the subset with minimum cardinality in $S^-(H)$. Then, $D^-(H) = K$.

**Proof.** We know from Proposition 1 that $|D^-(H)| \leq 1$. Then, ab absurdo suppose that $\forall L \in S^-(H), |K| \leq |L|$ and that $\exists W \in S^-(H) \mid (|W| > |K|) \wedge (D^-(H) = W)$. This means that $H \subseteq W \wedge H \subseteq K$ and by definition of NS-M $W \subseteq K \cap K \subseteq W$. If $W \subseteq K \Rightarrow |W| < |K|$; if $K \subseteq W \Rightarrow |K| < |W|$. So if $D^+(H) = W \Rightarrow |W| < |K|$.

The next proposition proves that the direct subsets of $H$ are always disjoints.

**Proposition 3** Let $C$ be a NS-C and $H \in C$ be a set, then $\forall K, L \in D^+(H), K \cap L = \emptyset$.

**Proof.** Ab absurdo suppose that $K \cap L \neq \emptyset \Rightarrow K \cap L = W$ such that $|W| \geq 1 \wedge W \nsubseteq K \wedge W \nsubseteq L \Rightarrow C$ is not a NS-C.
An Insight on Structural and Content Queries

When we deal with a NS-C, we can distinguish between structural and content queries. Structural queries ask for the relationships between the sets in a collection of subsets – e.g. “Return all the sets in $C$ which are supersets of the set $H \in C$”, in this case we are not interested in the actual content of the sets, instead we just want to know which sets are supersets of $H$. Content queries ask for the actual content (the elements) of the sets in a collection of subsets – e.g. “Return all the elements in $C$ belonging to supersets of $H \in C$” or “Return all the elements in $C$ belonging to subsets of $H \in C$”. A possible data structure to implement these queries has to take into account the structural and the content components of the NS-M. It is possible to propose a dictionary-based data structure which from the structural point-of-view, stores the inclusion dependencies between the sets; and, from the content point-of-view, it stores the materialization of the sets (i.e. the elements of each set).

If we consider a NS-C $C$ and a set $H \in C$, the structural query operations we point out are: DESCENDANTS($H$) which returns the sets which are descendants of $H$, ANCESTORS($H$) which returns the sets which are ancestors of $H$, CHILDREN($H$) which returns the sets which are children of $H$, and PARENT($H$) which returns the set which is the parent of $H$. The worst-case scenario for all these structural queries is represented by a NS-C structured as a chain. In the case of the DESCENDANTS($H$) query the worst-case input set $H$ is the top set because the query has to return all the sets in the given NS-C. All other structural query operations can be implemented to run in constant time for whichever collection of subsets and input set by exploiting the described properties of NS-M. For instance, the PARENT operation can be implemented in $O(1)$ time by exploiting the fact that every set in a NS-C has at most one direct superset as proved in Proposition 1.

The content query operations are: ELEMENTS($H$) which returns the elements in the set $H$, DESCENDANTELEMENTS($H$) which returns the elements in the descendants of $H$, ANCESTORELEMENTS($H$) which returns the elements in the ancestors of $H$, CHILDRENELEMENTS($H$) which returns the elements in the children of $H$, and PARENTELEMENTS($H$) which returns the elements in the parent of $H$.

The NS-M is built in such a way that each set contains all the elements of its descendants; therefore, it is possible to answer this query without browsing the collection of subsets (without browsing the hierarchy) or without passing through any structural query. We can see how these queries are independent with respect to their correspondent structural queries. In order to answer the content queries we do not need to browse the hierarchical structure of the collection of subsets.

Final Remarks and Future Works

By defining the NS-M we showed that it is possible to model hierarchies exploiting the inclusion property between sets in place of the binary relationship
between nodes adopted by the tree. We provided the theoretical basis for employing the NS-M as an alternative to the tree to model hierarchies, and for exploiting a set-theoretical environment for the definition of problems and solutions with hierarchies. It is possible to exploit the fact that content and structure components of hierarchies are modeled as independent building blocks of the NS-M to define structural and content queries. We gave a first insight from the theoretical point-of-view on how content queries do not strongly rely on structural queries as happens with the tree.

In the future we are going to conduct experiments on synthetic and real datasets to experimentally verify that both structural and content query operations are scalable and efficient in the NS-M. The experimental analysis will be aimed to show how the choice between the tree and the NS-M has to be made on an application basis by evaluating which operations are executed more frequently. It will be thus possible to investigate and propose data structures optimized for specific contexts and applications so that to gain further performances for the NS-M.

Acknowledgments

The author thanks Maristella Agosti and Nicola Ferro which supported this work and contributed to many of the presented ideas. The author would like to thank Carlo Meghini for his useful and accurate observations regarding the formalization and the properties of the NS-M. The PROMISE network of excellence\(^2\) (contract n. 258191) projects, as part of the 7th Framework Program of the European Commission, have partially supported the reported work.

References


\(^2\) [http://www.promise-noe.eu/](http://www.promise-noe.eu/)