

# Abstracting concepts from text documents by using a taxonomy

---

**E. Chernyak**<sup>1,4</sup>, O. Chugunova<sup>1</sup>, J. Askarova<sup>1</sup>, S. Nascimento<sup>2</sup>, B. Mirkin<sup>1,3</sup>

<sup>1</sup> Division of Applied Mathematics and Informatics, NRU-HSE, Moscow, Russia

<sup>2</sup> Department of Informatics, New University of Lisbon, Caparica, Portugal

<sup>3</sup> Department of Computer Science, Birkbeck University of London, London, UK

<sup>4</sup> Witology

# Contents

---

1. Statement of the problem
2. Method
3. Examples of application
4. Future work

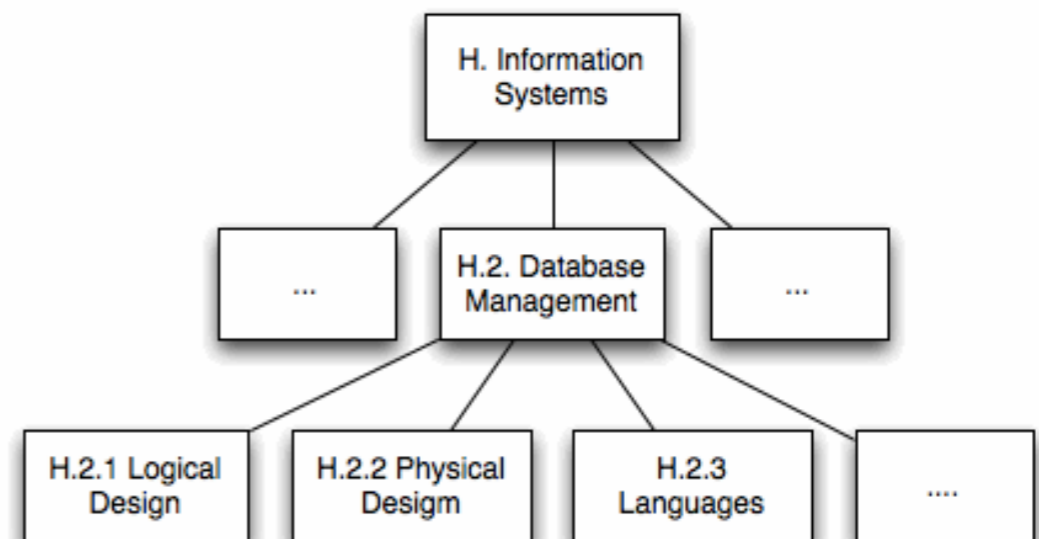
# Statement of the problem

---

- Interpretation of a text corpus over a taxonomy (the main part of an ontology)

Article: Two variable logic on data trees and XML reasoning, Journal of the ACM, 2003

Motivated by reasoning tasks for XML **languages**, the satisfiability problem of **logics** on data trees is investigated. The nodes of a data tree have a label from a finite set and a data value from a possibly infinite set. It is shown that satisfiability for two-variable first-order **logic** is decidable if the tree structure can be accessed only through the child and the next sibling predicates and the access to data values is restricted to equality tests. From this main result, decidability of satisfiability and containment for a data-aware fragment of XPath and of the implication problem for unary key and inclusion constraints is concluded. Motivated by reasoning tasks for XML **languages**, the satisfiability problem of **logics** on data trees is investigated. The nodes of a data tree have a label from a finite set and a data value from a possibly infinite set. It is shown that satisfiability for two-variable first-order **logic** is decidable if the tree structure can be accessed only through the child and the next sibling predicates and the access to data values is restricted to equality tests. From this main result, decidability of satisfiability and containment for a data-aware fragment of XPath and of the implication problem for unary key and inclusion constraints is concluded.



# Input

## Collection of the ACM Journal abstracts

## The ACM Computing Classification System (1998)

### Journal of the ACM (JACM)

Volume 56 Issue 3, May 2009

### Table of Contents


[← previous issue](#) | [next issue →](#)

[Introduction to PODS 2006 special section](#)

[Victor Vianu, Jan Van den Bussche](#)

Article No.: 11

doi>[10.1145/1516512.1516513](#)


Full text:  [PDF](#)

[Lower bounds for processing data with few random accesses to external memory](#)

[Martin Grohe, André Hernich, Nicole Schweikardt](#)

Article No.: 12

doi>[10.1145/1516512.1516514](#)

Full text:  [PDF](#)


We consider a scenario where we want to query a large dataset that is stored in e  
constrained resources in such a situation are the size of the main memory and th

[Two-variable logic on data trees and XML reasoning](#)

[Mikoaj Bojańczyk, Anca Muscholl, Thomas Schwentick, Luc Segoufin](#)

Article No.: 13

doi>[10.1145/1516512.1516515](#)

Full text:  [PDF](#)

...

- [D. Software](#)
  - [D.0 GENERAL](#)
  - [D.1 PROGRAMMING TECHNIQUES \(E\)](#)
    - [D.1.0 General](#)
    - [D.1.1 Applicative \(Functional\) Programming](#)
    - [D.1.2 Automatic Programming \(I.2.2\)](#)
    - [D.1.3 Concurrent Programming](#)
      - *Distributed programming*
      - *Parallel programming*

...

# Input

## Collection of the ACM Journal abstracts

## The ACM Computing Classification System (1998)

**Journal of the ACM (JACM)**  
Volume 56 Issue 3, May 2009

### Table of Contents


[← previous issue](#) | [next issue →](#)

[Introduction to PODS 2006 special section](#)

[Victor Vianu, Jan Van den Bussche](#)

Article No.: 11

doi>[10.1145/1516512.1516513](#)


Full text:  PDF

[Lower bounds for processing data with few random accesses to external memory](#)

[Martin Grohe, André Hernich, Nicole Schweikardt](#)

Article No.: 12

doi>[10.1145/1516512.1516514](#)

Full text:  PDF


We consider a scenario where we want to query a large dataset that is stored in e  
constrained resources in such a situation are the size of the main memory and th

[Two-variable logic on data trees and XML reasoning](#)

[Mikoaj Bojańczyk, Anca Muscholl, Thomas Schwentick, Luc Segoufin](#)

Article No.: 13

doi>[10.1145/1516512.1516515](#)

Full text:  PDF

...

- [D. Software](#)
  - [D.0 GENERAL](#)
  - [D.1 PROGRAMMING TECHNIQUES \(E\)](#)
    - [D.1.0 General](#)

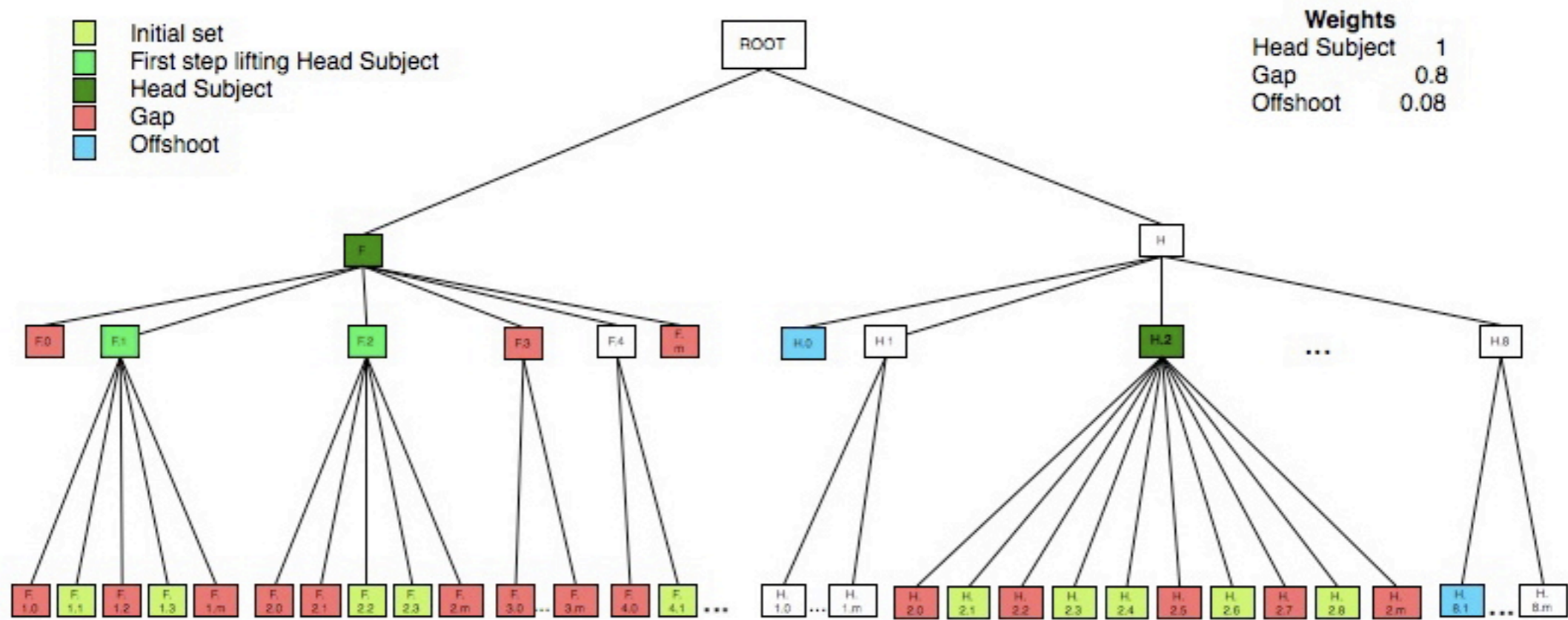
Primary Classification: F.1.1  
Additional Classification: F.1.3, H.2.4

- *Distributed programming*

Primary Classification: F.4.1  
Additional Classification: F.4.3, H.2.1,  
H.2.3, I.7.2

# Output

## Head subjects and related events (gap, offshoot)



### Profile of a text collection

Code	Membership value	ACM-CCS Topic
F.1.3	0.597	Complexity Measures and Classes
H.2.3	0.475	Languages
F.2.3	0.4009	Tradeoffs between Complexity Measures
H.2.1	0.3705	Logical Design
F.1.1	0.322	Models of Computation
H.2.4	0.2973	Systems
D.2.8	0.24	Metrics
H.2.8	0.2193	Database Applications
J.4	0.211	SOCIAL AND BEHAVIORAL SCIENCES
I.1.2	0.0178	Algorithms
...		

### Desired Interpretation

Head subjects:

H.2 DATABASE MANAGEMENT

F. Theory of Computation

# Method

---

## 1. Building a profile of the collection

- A. Annotated suffix tree for abstracts and keywords (Pampapathi, Mirkin, Levene, 2006)
- B. Scoring ACM-CCS leaves including references between them
- C. Clustering the profiles (if needed)

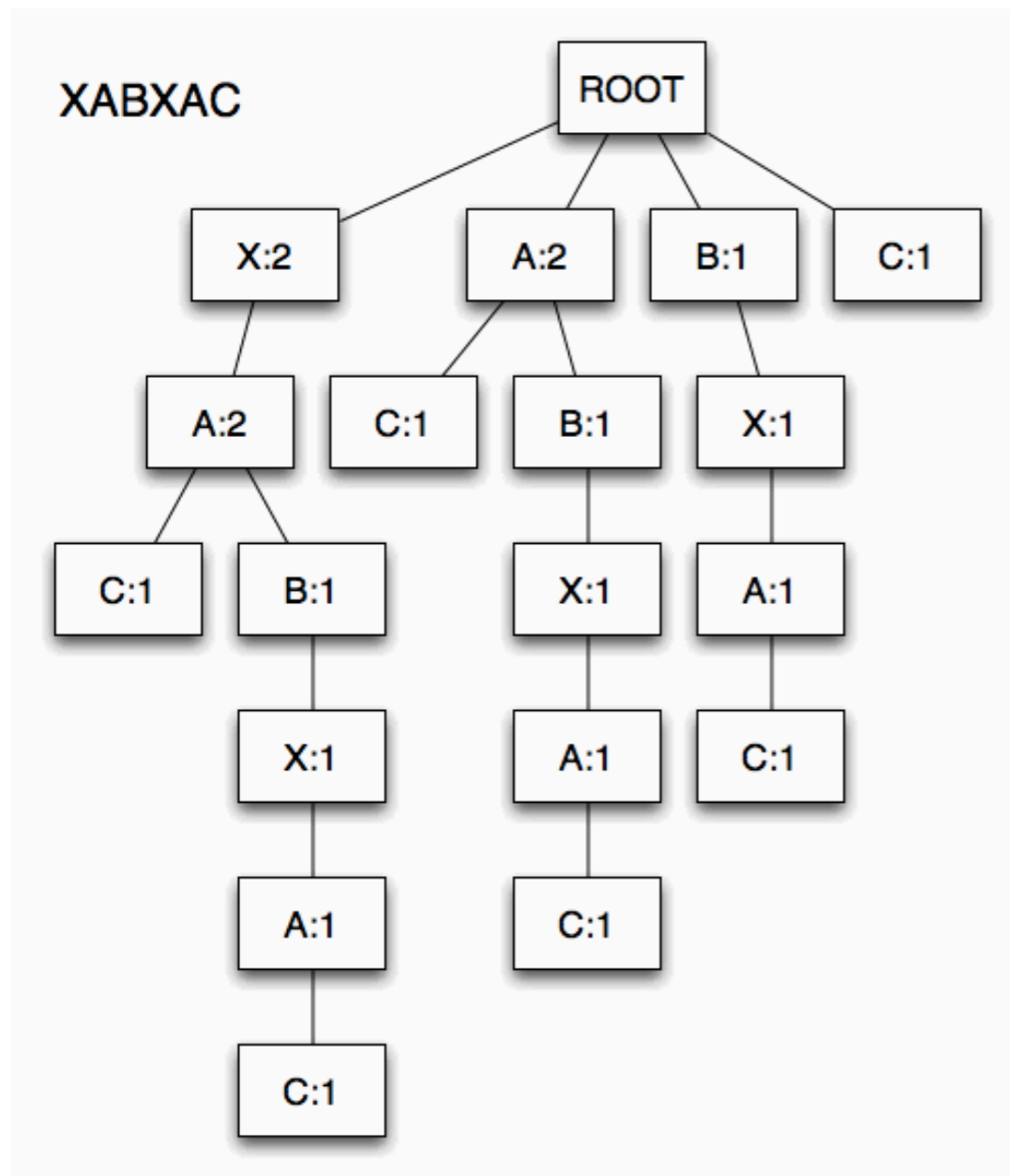
## 2. Lifting the profile in the taxonomy tree

- A. Specifying head subject, gap and offshoot penalty weights
- B. Parsimonious lifting (Mirkin, Nascimento, Fenner, Pereira, 2010)

# Annotated Suffix Tree (AST) for “xabxac”

---

- is used to compute and store the frequencies of all substrings of the string

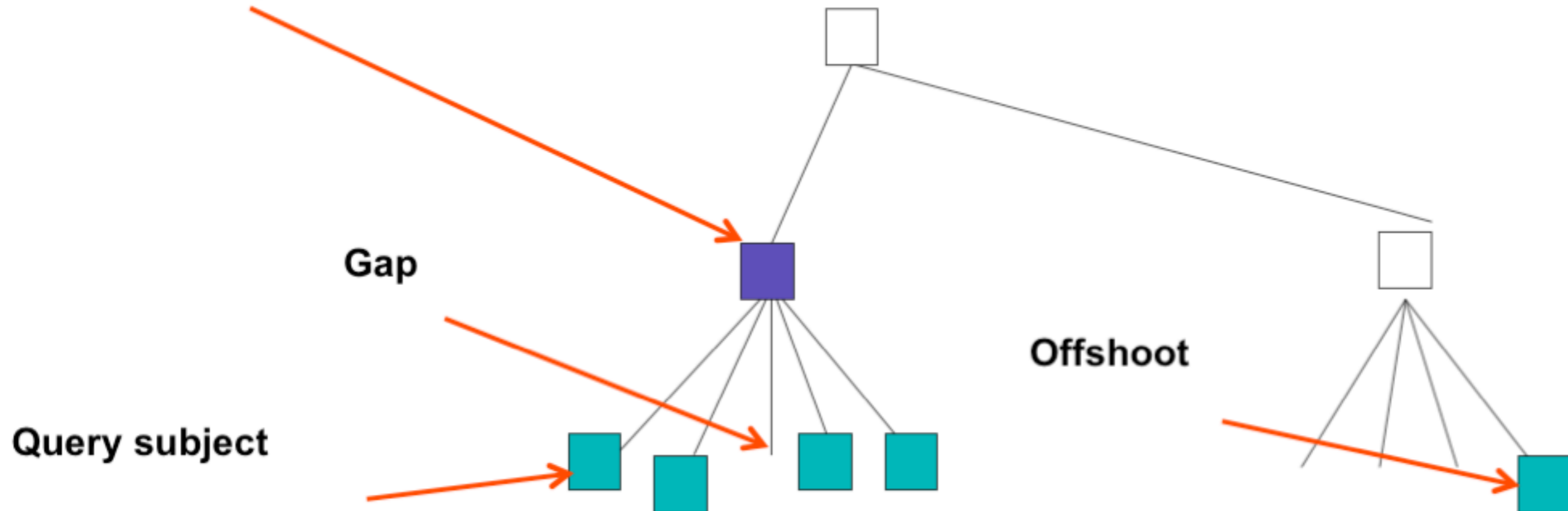




# Lifting

- Represent the thematic clusters in ACM-CCS **by** higher, more general, nodes depending on the inconsistencies (**Lift**)

**Head subject**



**Penalty  $H * \#Head\_Subj + G * \#Gap + O * \#Offshoot$**

# Two applications

---

- The Journal of ACM abstracts and the ACM-CCS
- Course syllabuses of Mathematics and Informatics disciplines and an in-house taxonomy of Mathematics and Informatics built using Supreme Attestation Committee of Russia documentation (in Russian)

# A “good” AST-profile

---

Article: Two variable logic on data trees and XML reasoning, Journal of the ACM, 2003

AST found profile			ACM-CCS index terms (manual annotation)		
ID	TE	ACM-CCS topic	ID	#	ACM-CCS topic
H.2.3	0.4541	Languages	H.2.3	0	Languages
I.1.3	0.4489	Languages and Systems	F.4.3	2	Formal Languages
F.4.3	0.3918	Formal Languages	H.2.1	12	Logical Design
D.4.5	0.3049	Reliability	F.4.1	27	Mathematical Logic
I.6.2	0.2578	Simulation Languages	I.7.2	52	Document Preparation

# A “poor” AST–profile

---

Article: Lower bounds for processing data with few random accesses to external memory.  
Journal of the ACM, 2003

AST found profile			ACM-CCS index terms (manual annotation)		
ID	TE	ACM–CCS topic	ID	#	ACM–CCS topic
H.2.8	0.4330	Database Applications	F.1.3	160	Complexity Measures and Classes
H.2.5	0.2904	Heterogeneous Databases	H.2.4	165	Systems
C.5.1	0.2630	Large and Medium (“Mainframe”) Computers	F.1.1	219	Models of Computation
J.1	0.2115	ADMINISTRATIVE DATA PROCESSING			
I.2.7	0.1870	Natural Language Processing			

# Conclusion

---

- Interpretation by producing profiles and lifting them in the taxonomy
- Issues
  - A. AST scoring – slow and noised
  - B. The taxonomies are not quite relevant
  - C. Penalty weights? (Future work: change the parsimony criterion for that of the maximum likelihood)
  - D. Assessment of the results