جامعة قطر
QATAR UNIVERSITY

Incremental Pseudo Rectangular Organization of Information Relative to a Domain

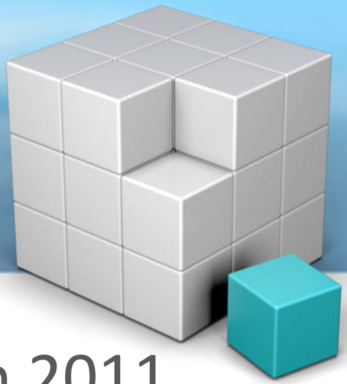
RAMiCS 13
Cambridge , UK
September 17 - 20

Sahar Ismail & Ali Jaoua

Agenda

1. Motivation and Objectives
2. Background
3. Solution Model
4. Evaluation
5. Applications / Demo
6. Conclusion

Motivation



- Information is doubling roughly every 11 hours in 2011

  Gartner and Accenture

- Worldwide information volume is growing at a minimum rate of 59% annually *IBM Study*

- 70% to 85% of data is "unstructured“ *Gartner*

- 87% of performance issues in application databases are related in some way to data growth *OAUG*

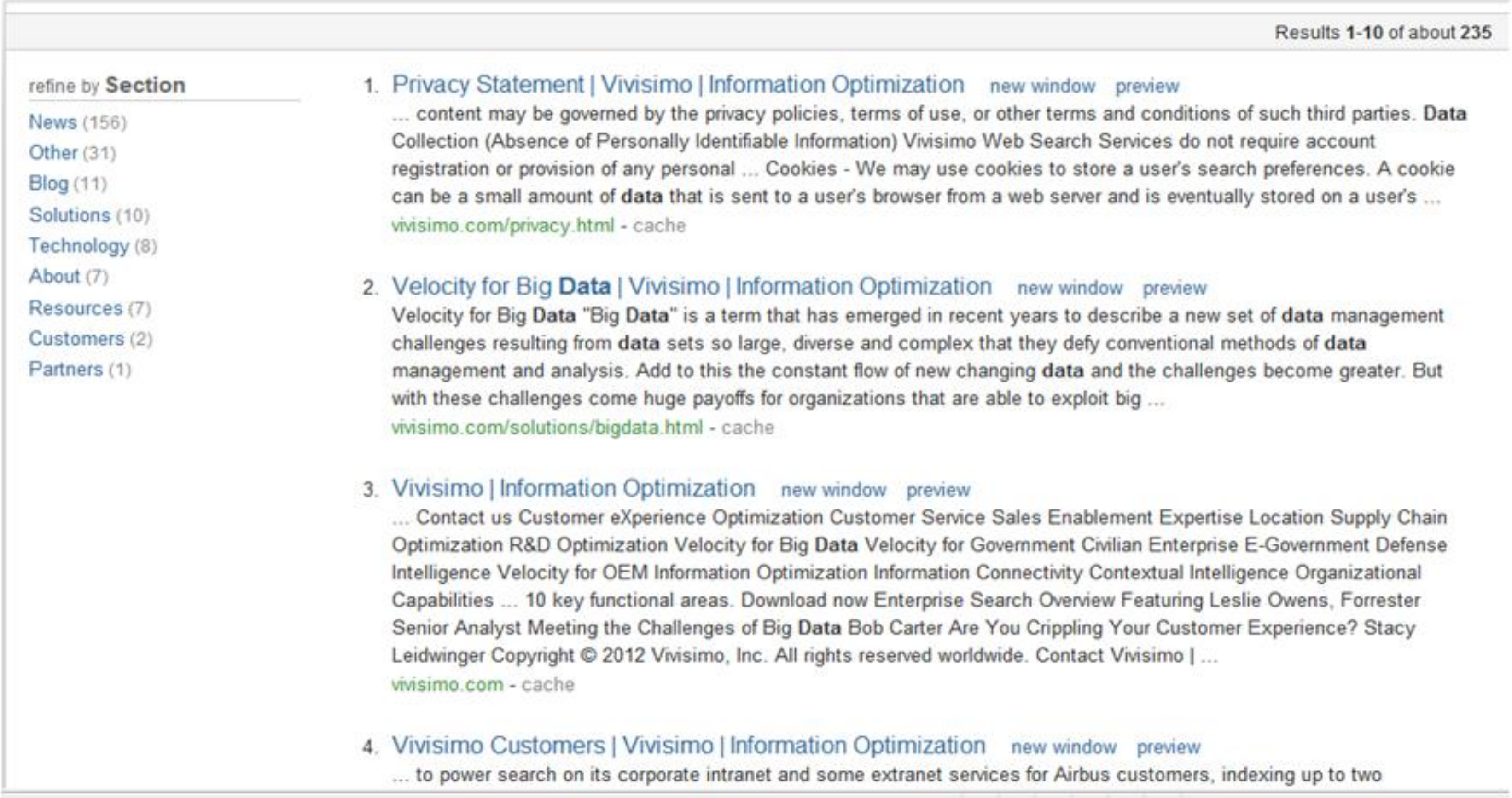
“.. There Is No Such Thing as Information Overload…No Giving Up, No Surrender ..”

M. Adrian

Motivation

Solutions in market: mostly organize info by date, section

→ Manually defined tags



Results **1-10** of about **235**

refine by **Section**

News (156)
Other (31)
Blog (11)
Solutions (10)
Technology (8)
About (7)
Resources (7)
Customers (2)
Partners (1)

1. Privacy Statement | Vivisimo | Information Optimization new window preview
... content may be governed by the privacy policies, terms of use, or other terms and conditions of such third parties. **Data** Collection (Absence of Personally Identifiable Information) Vivisimo Web Search Services do not require account registration or provision of any personal ... Cookies - We may use cookies to store a user's search preferences. A cookie can be a small amount of **data** that is sent to a user's browser from a web server and is eventually stored on a user's ...
vivisimo.com/privacy.html - cache

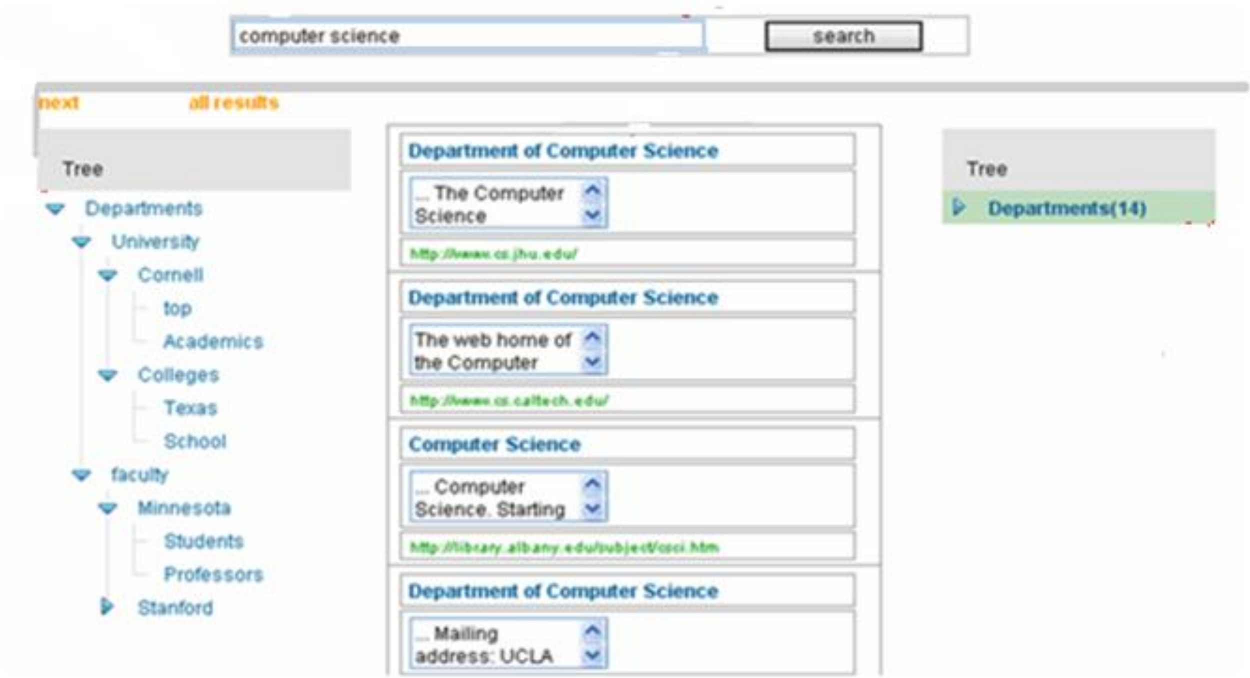
2. Velocity for Big **Data** | Vivisimo | Information Optimization new window preview
Velocity for Big **Data** "Big **Data**" is a term that has emerged in recent years to describe a new set of **data** management challenges resulting from **data** sets so large, diverse and complex that they defy conventional methods of **data** management and analysis. Add to this the constant flow of new changing **data** and the challenges become greater. But with these challenges come huge payoffs for organizations that are able to exploit big ...
vivisimo.com/solutions/bigdata.html - cache

3. Vivisimo | Information Optimization new window preview
... Contact us Customer eXperience Optimization Customer Service Sales Enablement Expertise Location Supply Chain Optimization R&D Optimization Velocity for Big **Data** Velocity for Government Civilian Enterprise E-Government Defense Intelligence Velocity for OEM Information Optimization Information Connectivity Contextual Intelligence Organizational Capabilities ... 10 key functional areas. Download now Enterprise Search Overview Featuring Leslie Owens, Forrester Senior Analyst Meeting the Challenges of Big **Data** Bob Carter Are You Crippling Your Customer Experience? Stacy Leidwinger Copyright © 2012 Vivisimo, Inc. All rights reserved worldwide. Contact Vivisimo | ...
vivisimo.com - cache

4. Vivisimo Customers | Vivisimo | Information Optimization new window preview
... to power search on its corporate intranet and some extranet services for Airbus customers, indexing up to two

Motivation

- Meta Search Engine → Too many clusters
- Not incremental, new website → do it all over again



- Departments
  - University
    - Cornell
      - top
      - Academics
    - Colleges
      - Texas
      - School
  - faculty
    - Minnesota
      - Students
      - Professors
    - Stanford

Department of Computer Science

... The Computer Science

http://www.cs.jhu.edu/

Department of Computer Science

The web home of the Computer

http://www.cs.caltech.edu/

Computer Science

... Computer Science. Starting

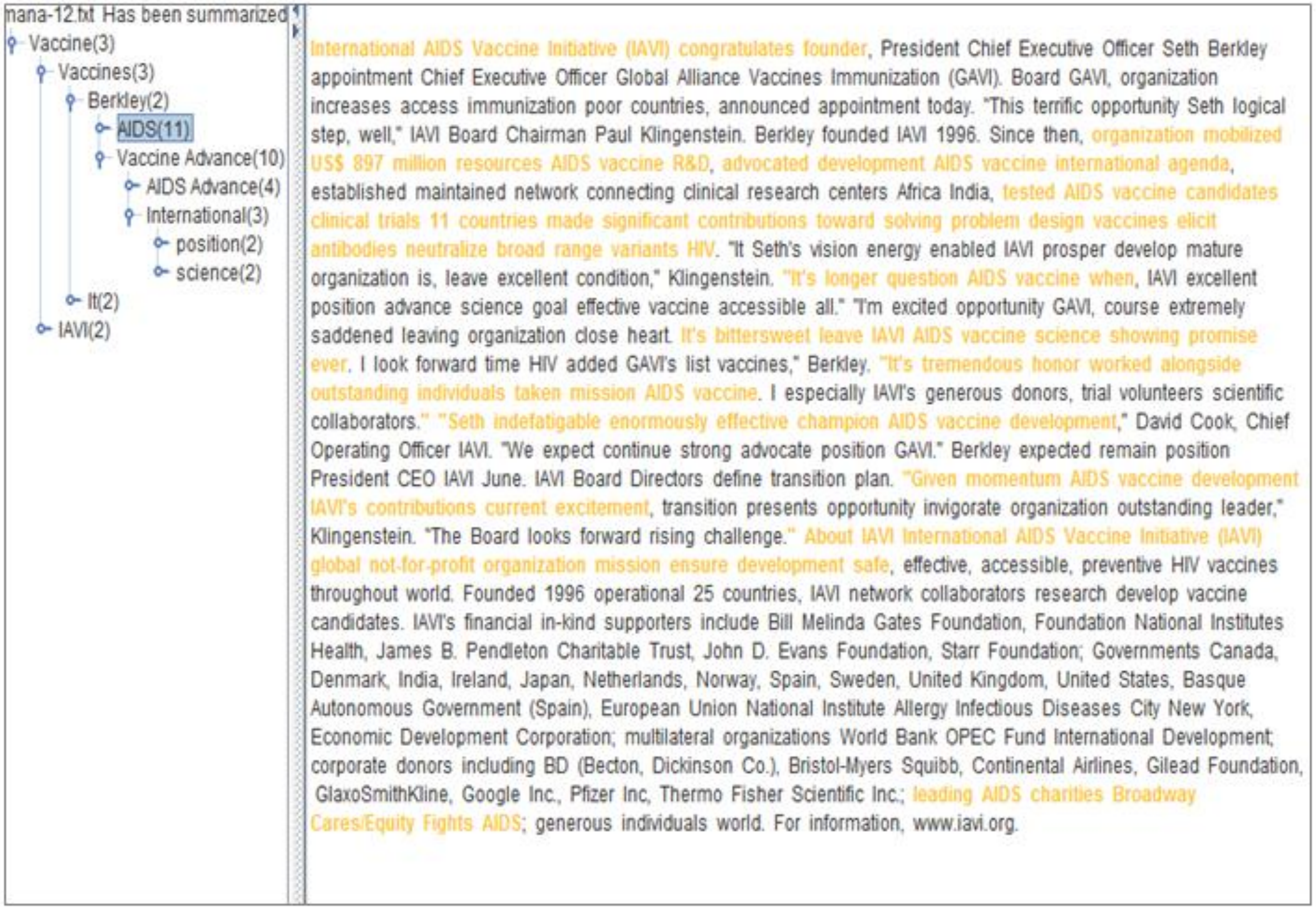
http://library.albany.edu/subject/csci.htm

Department of Computer Science

... Mailing address: UCLA

Motivation

- Text summarization → non incremental



nana-12.txt Has been summarized

- Vaccine(3)
  - Vaccines(3)
    - Berkley(2)
      - AIDS(11)
      - Vaccine Advance(10)
        - AIDS Advance(4)
        - International(3)
          - position(2)
          - science(2)
    - It(2)
  - IAVI(2)

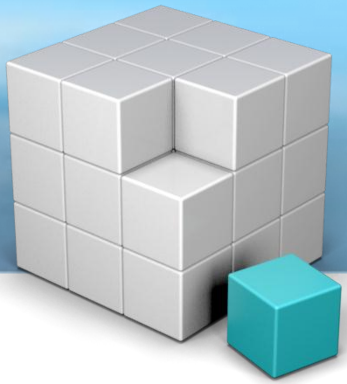
International AIDS Vaccine Initiative (IAVI) congratulates founder, President Chief Executive Officer Seth Berkley appointment Chief Executive Officer Global Alliance Vaccines Immunization (GAVI). Board GAVI, organization increases access immunization poor countries, announced appointment today. "This terrific opportunity Seth logical step, well," IAVI Board Chairman Paul Klingenstein. Berkley founded IAVI 1996. Since then, organization mobilized US$ 897 million resources AIDS vaccine R&D, advocated development AIDS vaccine international agenda, established maintained network connecting clinical research centers Africa India, tested AIDS vaccine candidates clinical trials 11 countries made significant contributions toward solving problem design vaccines elicit antibodies neutralize broad range variants HIV. "It Seth's vision energy enabled IAVI prosper develop mature organization is, leave excellent condition," Klingenstein. "It's longer question AIDS vaccine when, IAVI excellent position advance science goal effective vaccine accessible all." "I'm excited opportunity GAVI, course extremely saddened leaving organization close heart. It's bittersweet leave IAVI AIDS vaccine science showing promise ever. I look forward time HIV added GAVI's list vaccines," Berkley. "It's tremendous honor worked alongside outstanding individuals taken mission AIDS vaccine. I especially IAVI's generous donors, trial volunteers scientific collaborators." "Seth indefatigable enormously effective champion AIDS vaccine development," David Cook, Chief Operating Officer IAVI. "We expect continue strong advocate position GAVI." Berkley expected remain position President CEO IAVI June. IAVI Board Directors define transition plan. "Given momentum AIDS vaccine development IAVI's contributions current excitement, transition presents opportunity invigorate organization outstanding leader," Klingenstein. "The Board looks forward rising challenge." About IAVI International AIDS Vaccine Initiative (IAVI) global not-for-profit organization mission ensure development safe, effective, accessible, preventive HIV vaccines throughout world. Founded 1996 operational 25 countries, IAVI network collaborators research develop vaccine candidates. IAVI's financial in-kind supporters include Bill Melinda Gates Foundation, Foundation National Institutes Health, James B. Pendleton Charitable Trust, John D. Evans Foundation, Starr Foundation; Governments Canada, Denmark, India, Ireland, Japan, Netherlands, Norway, Spain, Sweden, United Kingdom, United States, Basque Autonomous Government (Spain), European Union National Institute Allergy Infectious Diseases City New York, Economic Development Corporation; multilateral organizations World Bank OPEC Fund International Development; corporate donors including BD (Becton, Dickinson Co.), Bristol-Myers Squibb, Continental Airlines, Gilead Foundation, GlaxoSmithKline, Google Inc., Pfizer Inc, Thermo Fisher Scientific Inc.; leading AIDS charities Broadway Cares/Equity Fights AIDS; generous individuals world. For information, www.iavi.org.

Incremental Pseudo Structuring System IPS

- **Objectives:**
  - Perform incremental information organization using pseudo maximal rectangles (new conceptual constructs)
  - Process large scale data incrementally with reasonable quality and efficiency
  - Produce domain-biased information structuring

- **Supported applications:**
  - Macro-level structuring (Documents X Words)
  - Text Summarization (Sentences X Words)
  - Feature extraction

Background
Information Representation(1/2)



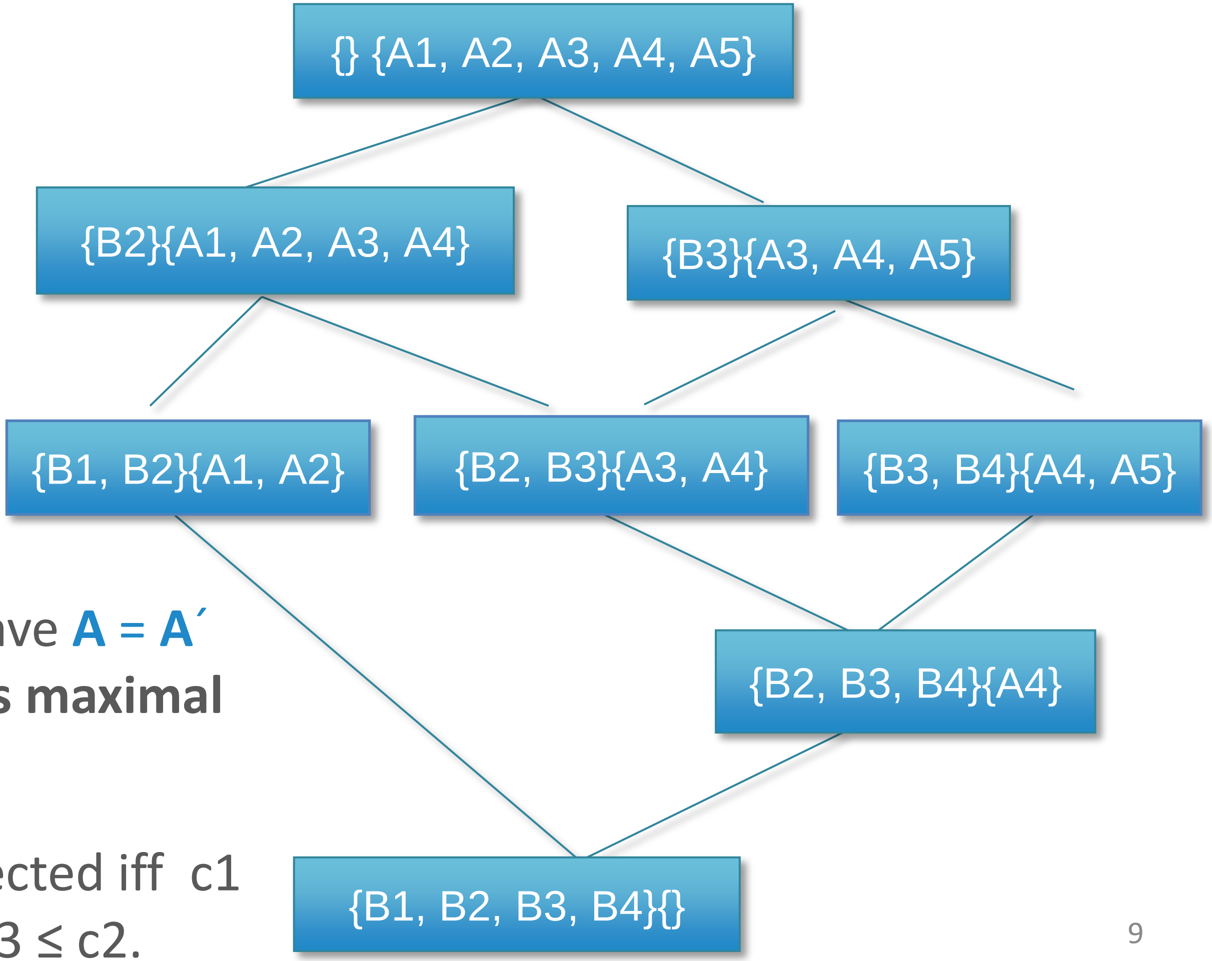
Formal Context I: (A,B,R)

R	B1	B2	B3	B4
A1	1	1	0	0
A2	1	1	0	0
A3	0	1	1	0
A4	0	1	1	1
A5	0	0	1	1

If **AXB** ⊆ **A´XB´**⊆ **R**, and we have **A = A´** and **B = B´** ➔ the **Rectangle is maximal (Concept)**

Concepts c1 and c2 are connected iff c1 ≤ c2 and ∄ c3 such that c1 ≤ c3 ≤ c2.

Galois Lattice



{} {A1, A2, A3, A4, A5}

{B2}{A1, A2, A3, A4}

{B3}{A3, A4, A5}

{B1, B2}{A1, A2}

{B2, B3}{A3, A4}

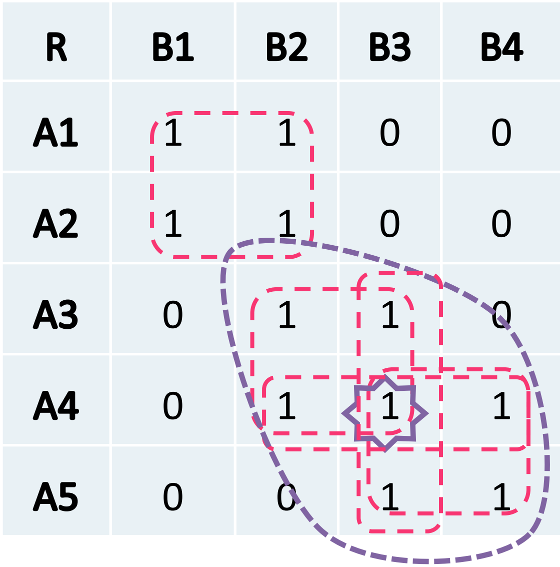
{B3, B4}{A4, A5}

{B2, B3, B4}{A4}

{B1, B2, B3, B4}{}

Background
Information Representation(2/2)

Formal Context I: (A,B,R)



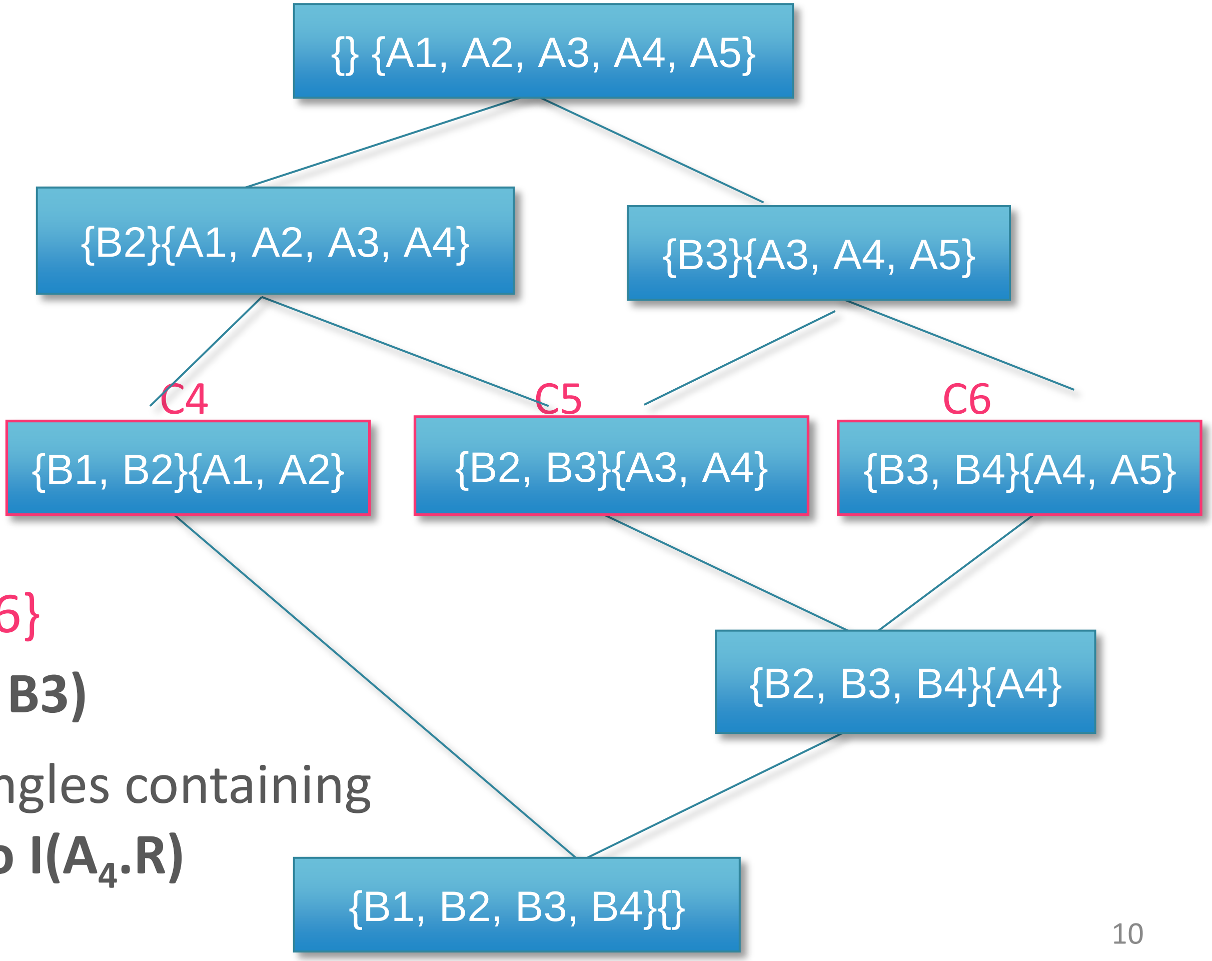
R	B1	B2	B3	B4
A1	1	1	0	0
A2	1	1	0	0
A3	0	1	1	0
A4	0	1	1	1
A5	0	0	1	1

Minimal Coverage: {C4,C5,C6}

Pseudo Max. Rectangle (A4, B3)

Is the union of all max. rectangles containing (A_4,B_3): PS = **I(B_3.R^{-1}) o R o I(A_4.R)**

Lattice of Context I

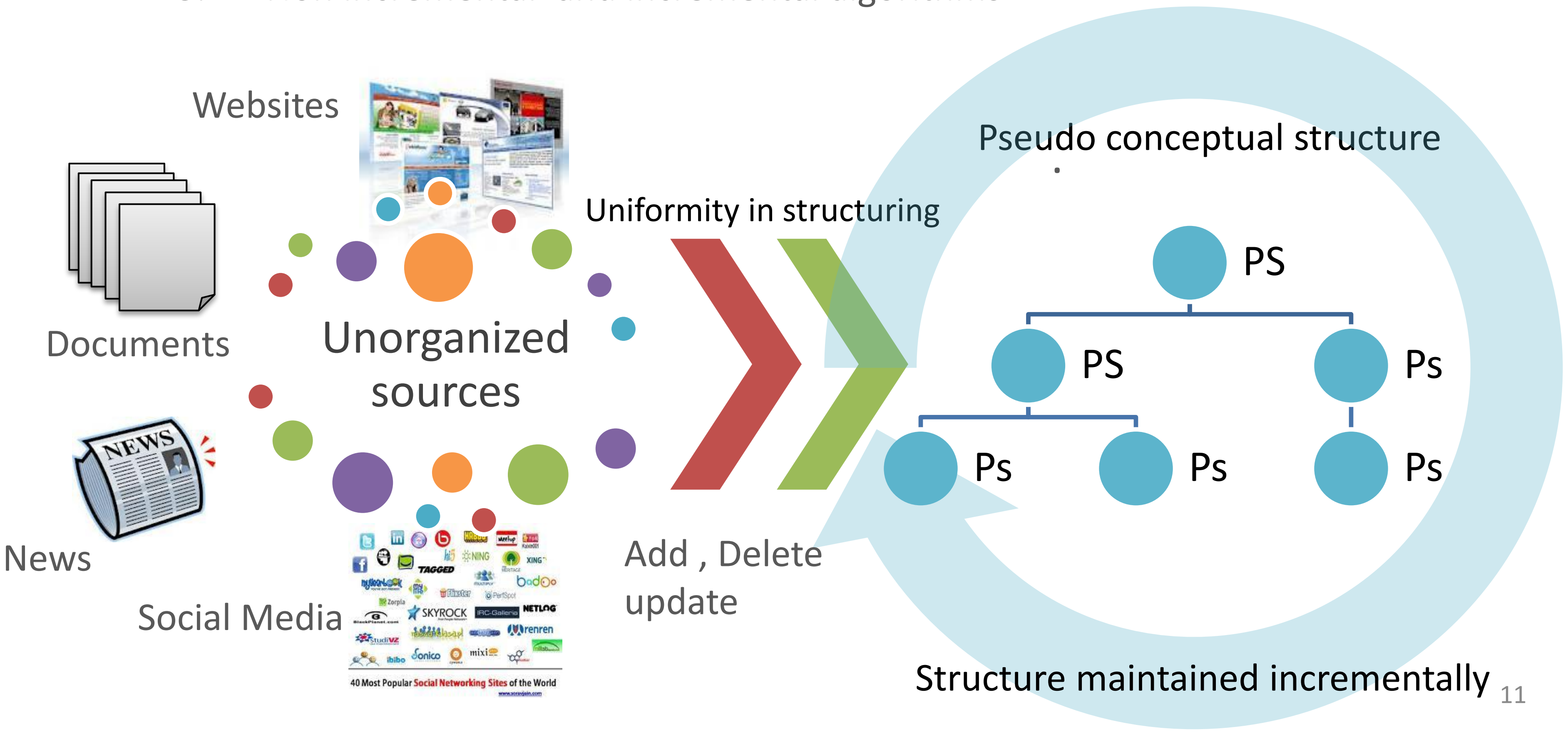


Solution Model

Overview

The solution was modeled in terms of the following dimensions:

1. Domain of knowledge
2. Information store
3. Non incremental and incremental algorithms



Websites
Documents
Unorganized sources
News
Social Media
40 Most Popular Social Networking Sites of the World
Uniformity in structuring
Add , Delete update
Pseudo conceptual structure
PS
PS
Ps
Ps
Ps
Ps
Structure maintained incrementally

Solution Model
Domain of knowledge

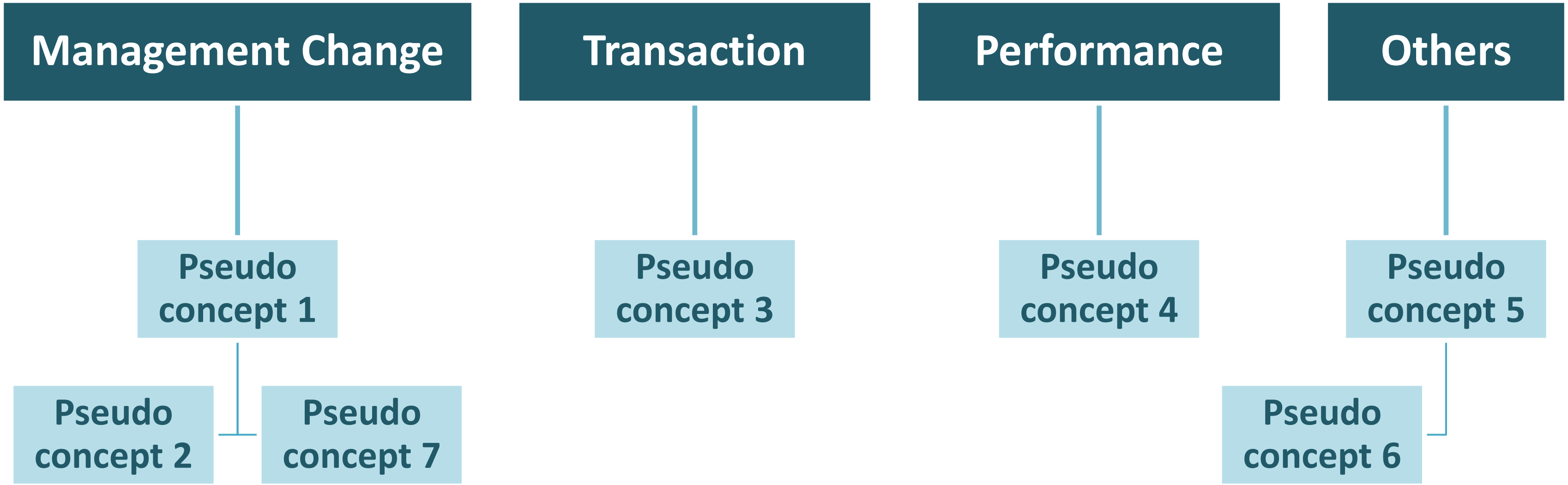
- Domain information categorized into disjoint bag of words
- Each bag represents a category and contains representative category words
- Aid labeling and structuring → more semantics but less minimality

Management Change	• retire, appoint, resign, assign, promote, demote, hire
Transaction	• buy, sell, Lease, rent, deal, loan, contract, asset, exchange
Performance	• grow, shrink, increase, decrease, lose , gain , profit, earn, drop, jump, discount, raise
Others	• <Empty bag>

FWATCH Domain Bags of words

Solution Model
Structured information store

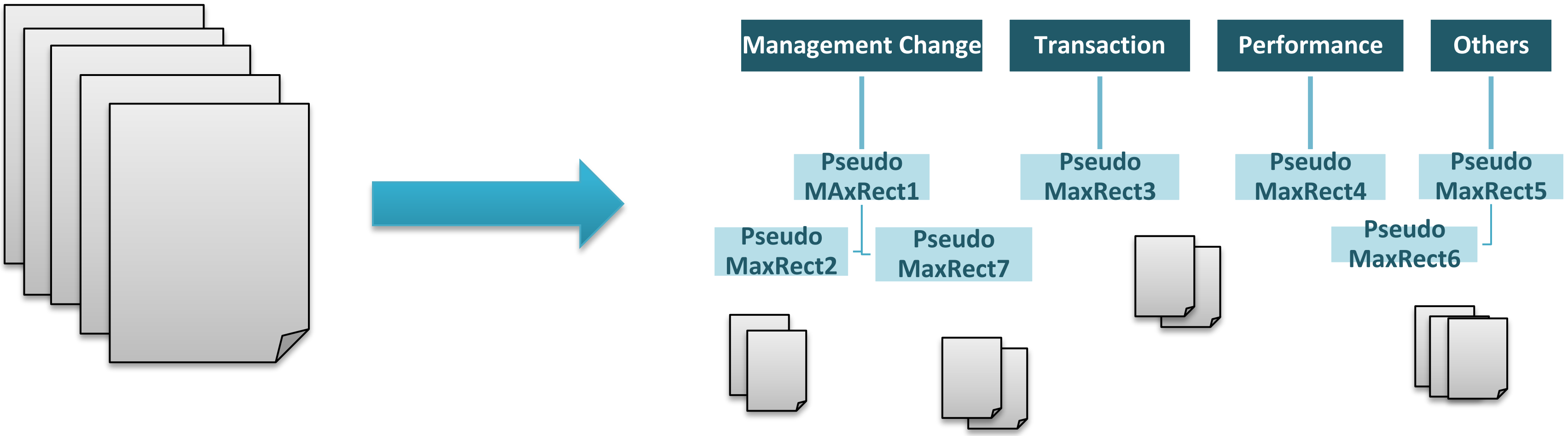
- Input: Binary relation representing documents
- Output of the structuring process:
  - Multi-root tree with category nodes at the first level
  - Subsequent levels under each category hold pseudo max. rectangles organized as a heap



Solution Model

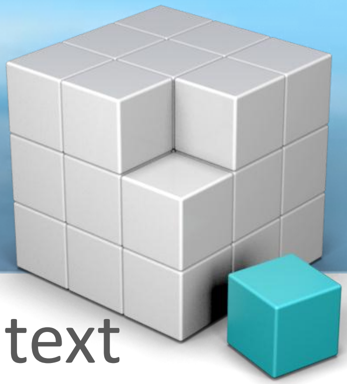
Algorithms

- ✓ Find minimal pseudo rectangular coverage non-incrementally + structure relevant to domain
- ✓ Maintain the structure incrementally in acceptable **quality** and reasonable **efficiency**



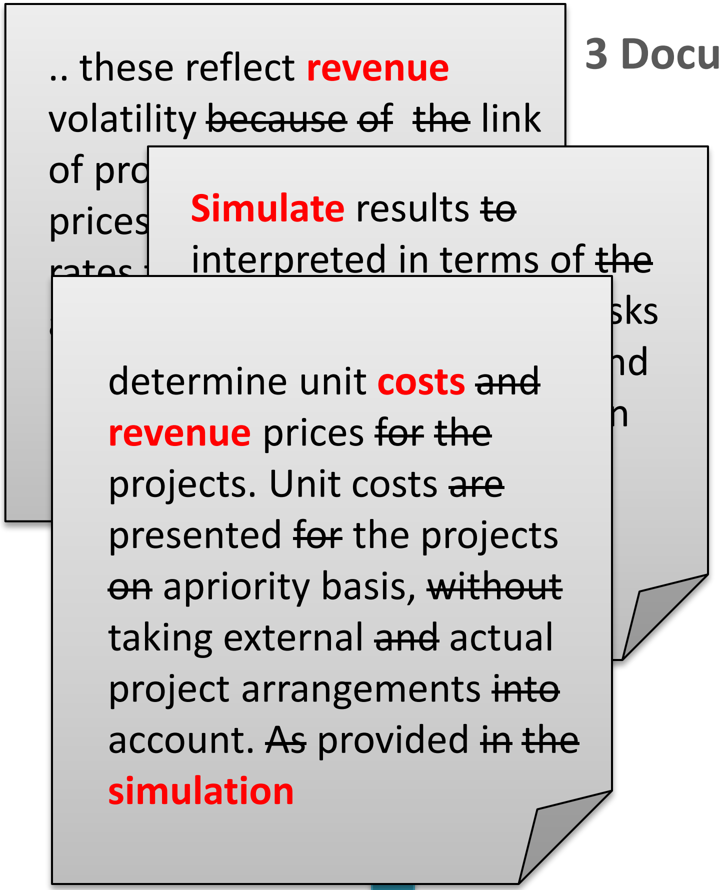
Methodology

Non-incremental algorithm (1/3)



1. Pre-preprocessing + NLP → Create Formal Context

3 Documents



.. these reflect **revenue** volatility ~~because of~~ ~~the~~ link of pro
prices
rates

Simulate results ~~to~~ interpreted in terms of ~~the~~
sks
nd
n

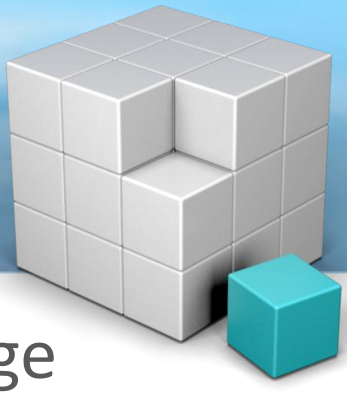
determine unit **costs** ~~and~~ **revenue** prices ~~for the~~ projects. Unit costs ~~are~~ presented ~~for~~ the projects ~~on~~ apriority basis, ~~without~~ taking external ~~and~~ actual project arrangements ~~into~~ account. ~~As~~ provided ~~in the~~ **simulation**

Pair	weight
(1,1)	2
(2,2)	3
(3,1)	5
(3,2)	1
…	…
(3,n)	-1

R[3][n]: 3 documents associated to n non-empty words

R	revenue	simulate	…	cost
doc1	1	0	…	0
doc2	0	1	…	0
doc3	1	1	..	1

Methodology
Non-incremental algorithm (2/3)



2. Calculate minimal pseudo-conceptual coverage

2.1 Sort pairs in descending order of their weights

Pair	weight
~~(3,1)~~	~~5~~
~~(2,2)~~	~~3~~
~~(1,1)~~	~~2~~
~~(3,2)~~	~~1~~
…	…
~~(3,n)~~	~~1~~

2.2 Highest weight pair (Familiar with respect to domain)

2.6 Next uncovered highest weight pair

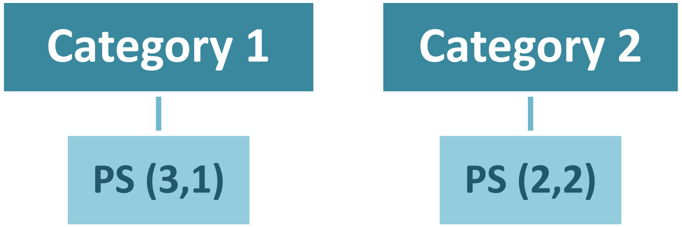
← **2.4** Mark all pairs as covered

← **2.8** Mark all pairs as covered

2.3 find
PS (3,1) =I(1.R^{-1}) o R o I(3.R)
= {(3,1),(3,2), (3,n) , (1,1)}

2.7 PS (1,1) =I(1.R^{-1}) o R o I(1.R)
= {(1,3),(1,1)}

2.5 Select category and label

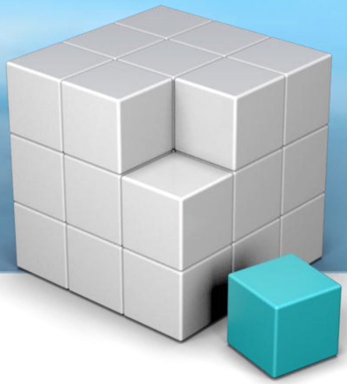


Category is the most overlapping with ps words → label is the best not selected

Stop when all pairs are covered

Methodology

Non-incremental algorithm (3/3)



- **High degree of scalability**
  - Number of pseudo-concepts is small compared to size of concepts in a lattice or a minimal coverage set (exponential)
  - Number of pseudo concepts is bound by the number of pairs in the relation
- **Efficient**
  - Worst case time complexity is in **O (N^2)** ; N is density of 1’s in R
  - Best case time complexity is in **O(N log N)**

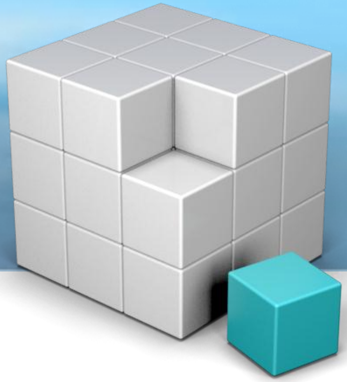
- Resulting structure is **domain sensitive**

Solution Model
Incremental Structuring Algorithm

The new methods cover the following update cases:

- Addition and deletion of domain elements
- Addition and deletion of codomain elements
- Addition and deletion of associations
- Updates to the domain of knowledge

Solution Model
Incremental Add Algorithm (1/4)

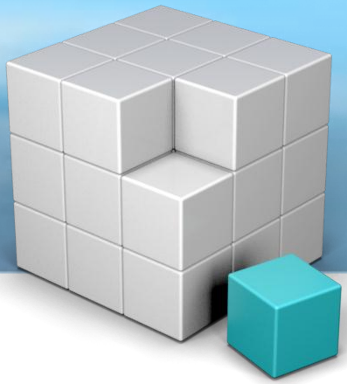


Algorithmic skeleton

1. Identify parts of R to update on the addition of a pair or a domain or codomain element
2. Identify pseudo maximal rectangles to be updated with the new information
3. Structure information identified as irrelevant to all existing pseudo maximal rectangles in a similar fashion to the non-incremental algorithm

Methodology

Incremental Add Algorithm (2/4)



1. What to update in R

On the arrival of new information , some pairs will require weight recalculation

$$\mathbf{s}(\mathbf{d},\boldsymbol{w}) = \left(|\mathbf{d}.\boldsymbol{R}| \times \left|\mathbf{w}.\boldsymbol{R}^{-1}\right|\right) - \left(|\mathbf{d}.\boldsymbol{R}| + \left|\mathbf{w}.\boldsymbol{R}^{-1}\right|\right)$$

→ Overhead increases as the cardinality of the update set increase

New domain element

	a	b	c
1	1	1	0
2	1	1	0
3	1	0	0
4	0	1	1
5	1	0	1
x	1	1	0

New codomain element

	a	b	c	y
1	1	1	0	1
2	1	1	0	0
3	1	0	0	0
4	0	1	1	1
5	1	0	1	0

New Pair

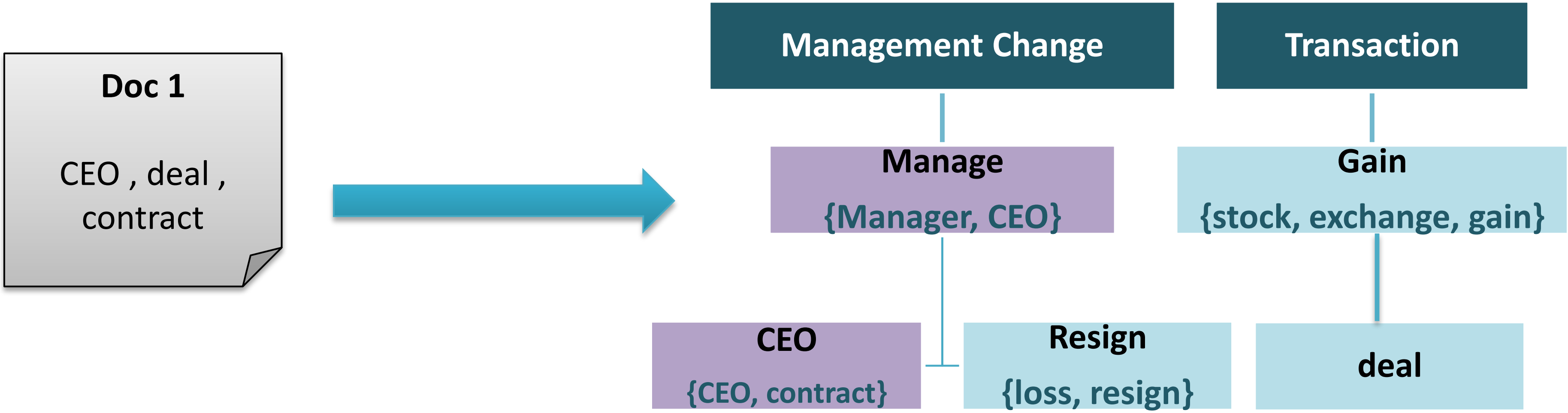
	a	b	c
1	1	1	0
2	1	1	0
3	1	0	0
4	0	**1**	1
5	1	0	1

Methodology
Incremental Add Algorithm (3/4)

2. What to update in PS tree

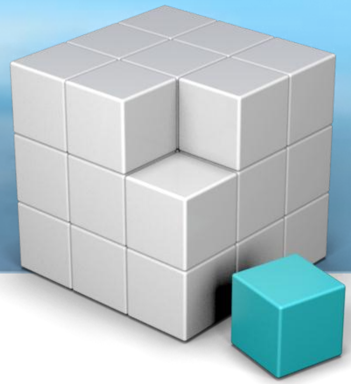
Relevant pairs → Find home from existing pseudo max. rectangles

Not relevant pairs → Structure as new



Methodology

Incremental Add Algorithm (4/4)

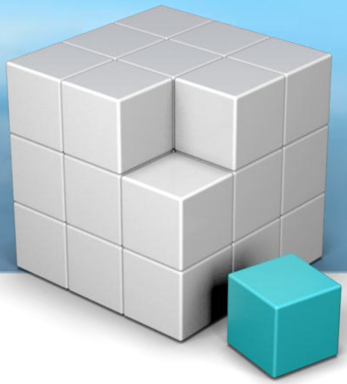


- Different addition scenarios:
  - **O (n log n)** in the best case and **O (n2 + N log N)** in the worst case ; (n : count of new pairs, N: count of all pairs in R)

- Add algorithms perform better when the Context is stable or n << N ; (N is count of all pairs in R)

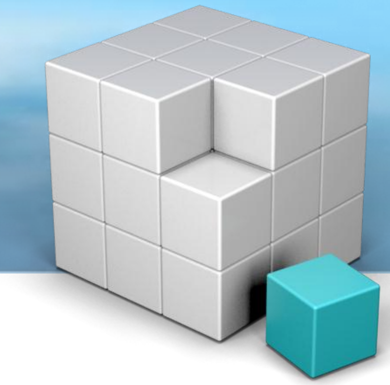
- Weight recalculation → Big computational overhead
  - Relax the immediate weight update condition
  - Drawback: Info store isn't up to date at all times (update strategy)

Evaluation
Approach and metrics(1/3)



- Incremental vs. non-incremental methods
- Metrics:
  - Cohesion: measures inner tightness in a pseudo concept
  - Separation: measures inter-pseudo concept overlapping
  - and performance (incremental vs. non incremental)
- Testing data: ~1020 documents from the NSF Research Award Abstracts 1990-2003
- Bag-of-words used represent the entire vocabulary of the documents in use

Evaluation
Approach and metrics(2/3)



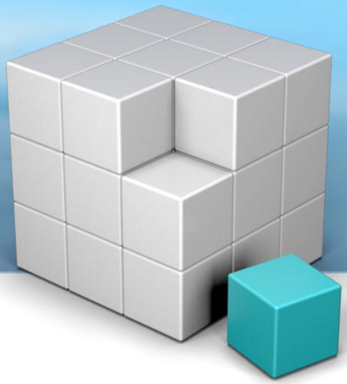
- Quality of structure – **Cohesion**
  - Measures the degree of tightness within a pseudo-concept (how related the contained pairs are)

$$\mathrm{Cohesion} = \sum_{i=1}^{k} \mathrm{Cohesion(PSi)}$$

$$\mathrm{Cohesion(PSi)} = \frac{|\mathrm{PSi}|^2}{|\mathrm{Ai}| \times |\mathrm{Bi}|}$$

- High cohesion is favorable , higher inner tightness means
  - Less pseudo maximal rectangles to maintain
  - Better scalability
  - Perfect cohesion → Pseudo maximal rectangle is a maximal rectangle

Evaluation
Approach and metrics(3/3)



- Quality of structure – **Separation**
  - Measures how independent or distinct two pseudo maximal rectangles are

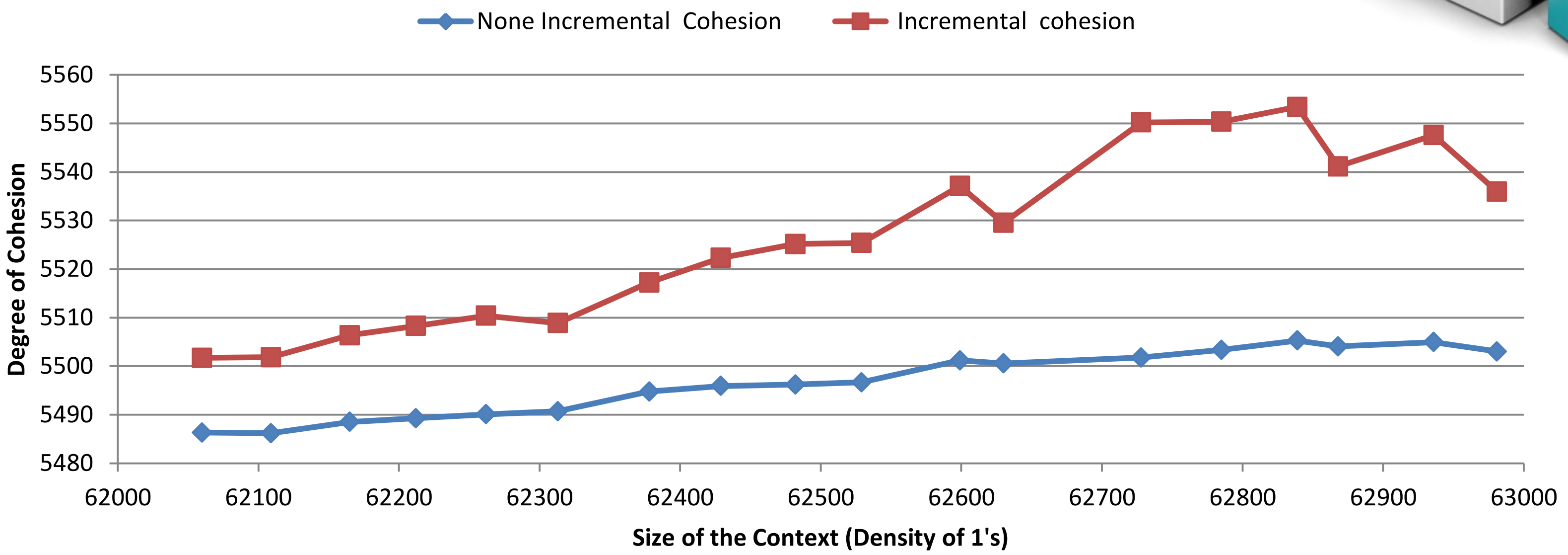
$$\mathrm{Separation\ (PSi, PSj)} = \frac{1}{|\mathrm{A_i} \ \cap \ \mathrm{Aj}\,| \times \ |\mathrm{B_i} \ \cap \ \mathrm{Bj}\,|}$$

$$\mathrm{Separation} \ = \sum_{\mathrm{j=0}}^{\mathrm{k-1}} \sum_{\mathrm{j=i+1}}^{\mathrm{k}} \mathrm{Separation\ (PSi, PSj)}$$

- Separation is favorable for incremental management, less overlapping means:
  - less update requirements in case of change
  - more stable structure

Evaluation – Incremental Add

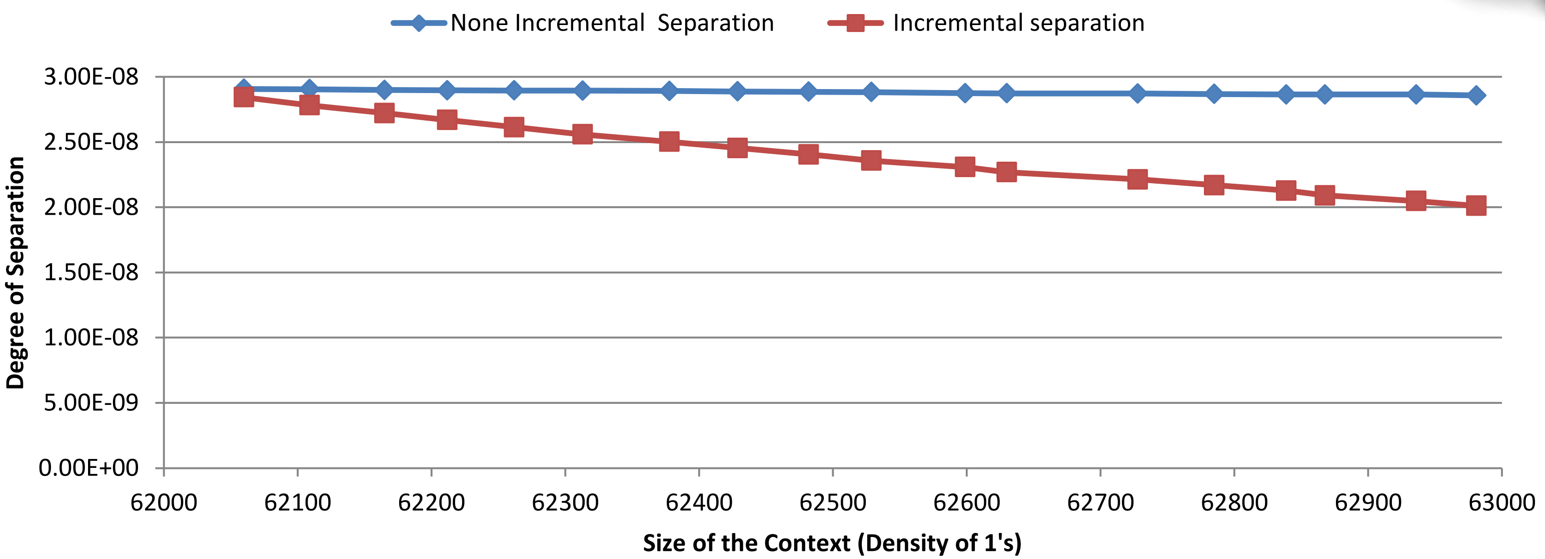
Experimentation results - Cohesion



- Cohesion is increasing as more documents are added incrementally
- Incremental-add algorithm favors updating existing relevant pseudo concepts over creating new pseudo concepts
- Relevant pairs are added which increases the degree of completeness of pseudo concept and increase its size

Evaluation – Incremental Add

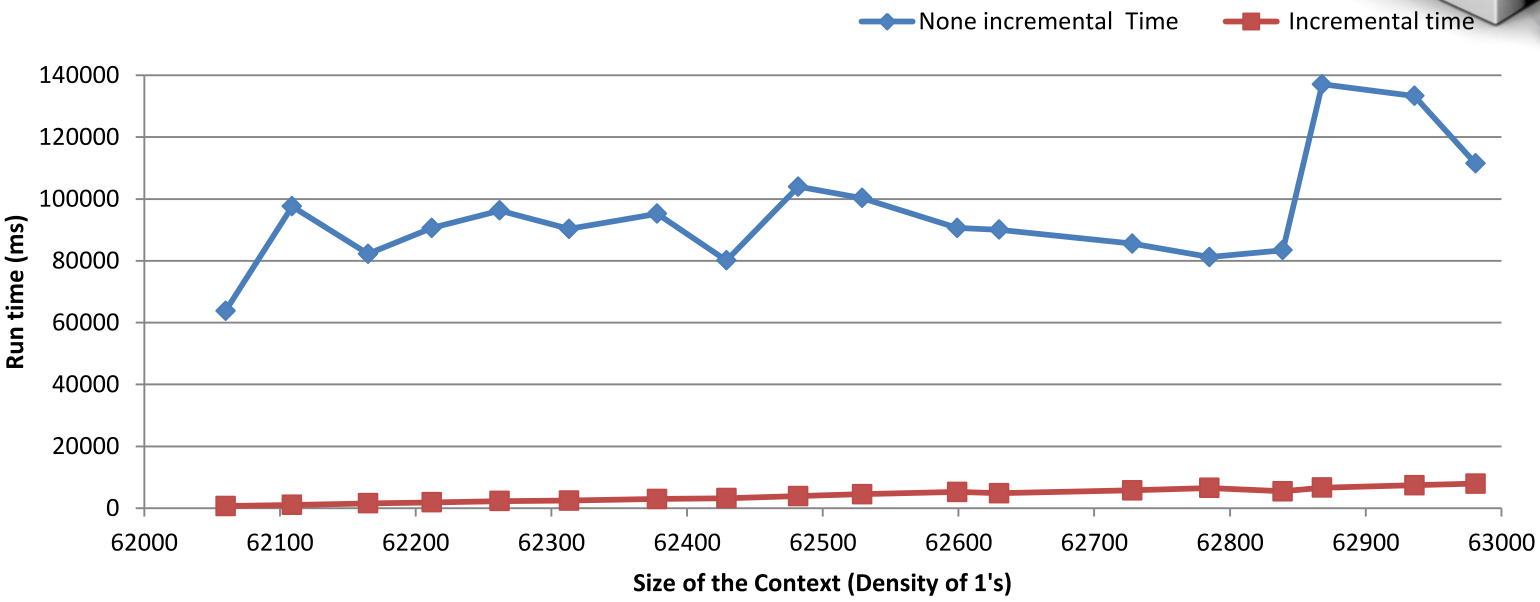
Experimentation results - Separation



- Separation is declining compared to the non-incremental method as more documents are added to the context
- New documents added update all relevant pseudo concepts which decreases separation → workaround: Don’t update all

Evaluation – Incremental Add

Experimentation results - Performance



- Time difference between the two approaches is considerable
- Both grow as a function of the density of 1’s in the context
- Counting the accumulative time for obtaining a structure using the incremental method shows it isn’t suited for structuring a space from scratch

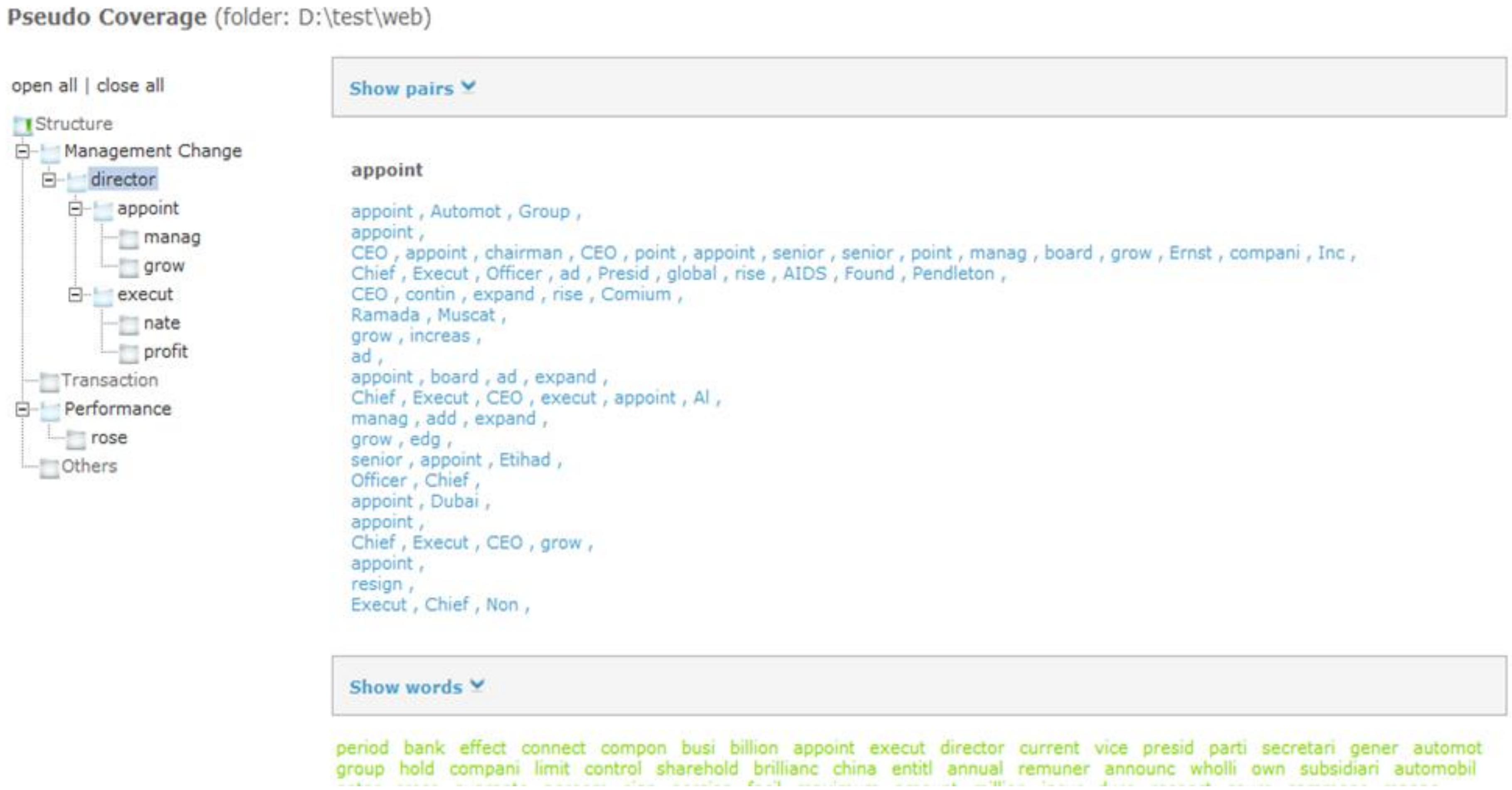
Conclusion

- IPS System uses new methods to handle incremental changes in an information store
- New conceptual constructs were used → Pseudo maximal rectangles
- Minimal pseudo coverage provides high level summary → promising scalability.
- Domain information is used to reduce the linguistic noise and improve labeling process
- System can handle incremental corpus organization, text summarization and feature-extraction

Applications

Macro-level structuring (Documents X Words)

Macro structuring and Feature Extraction



Pseudo Coverage (folder: D:\test\web)

open all | close all

- Structure
  - Management Change
    - director
      - appoint
        - manag
        - grow
      - execut
        - nate
        - profit
  - Transaction
  - Performance
    - rose
  - Others

Show pairs

appoint

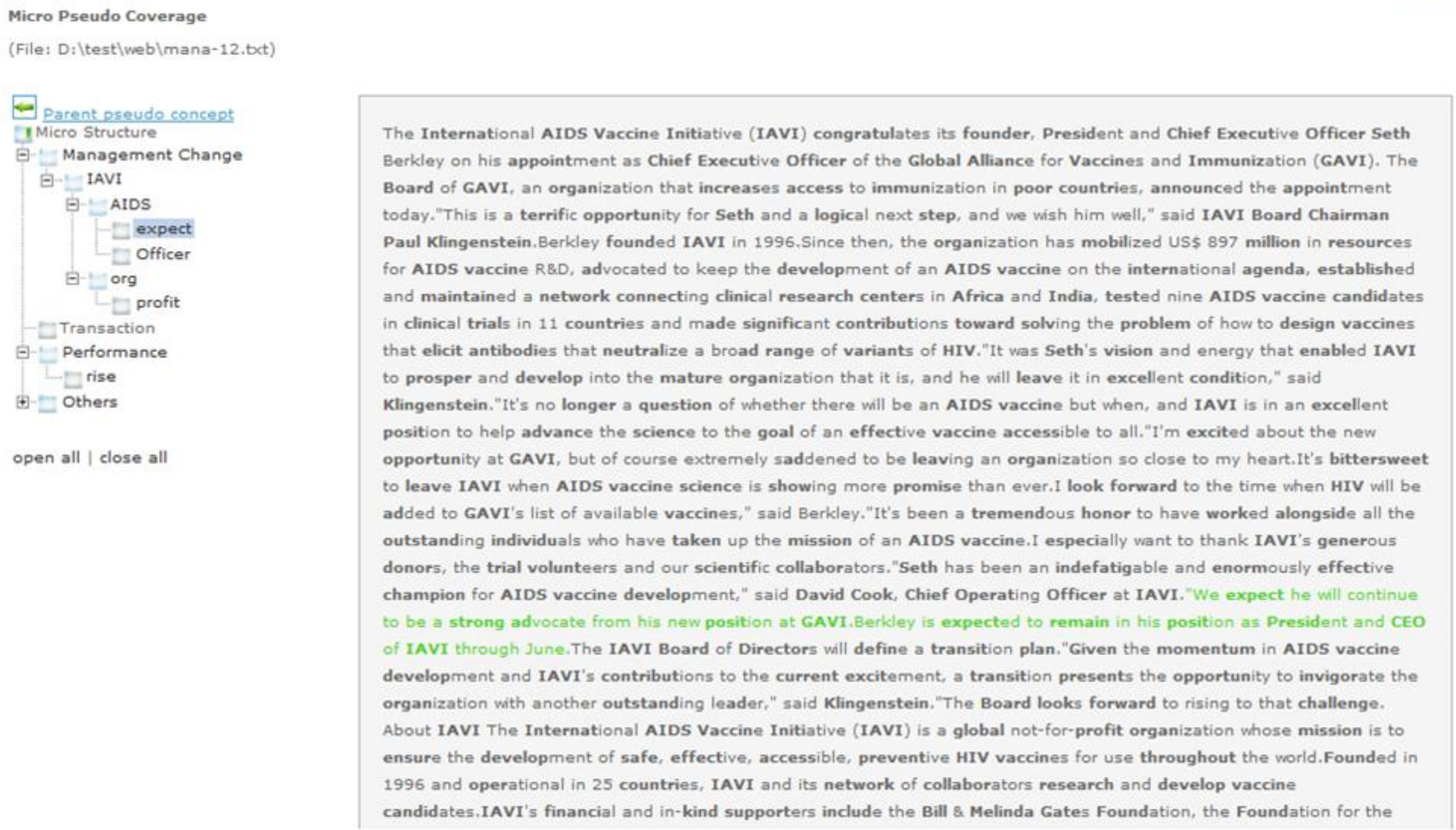
appoint , Automot , Group ,
appoint ,
CEO , appoint , chairman , CEO , point , appoint , senior , senior , point , manag , board , grow , Ernst , compani , Inc ,
Chief , Execut , Officer , ad , Presid , global , rise , AIDS , Found , Pendleton ,
CEO , contin , expand , rise , Comium ,
Ramada , Muscat ,
grow , increas ,
ad ,
appoint , board , ad , expand ,
Chief , Execut , CEO , execut , appoint , Al ,
manag , add , expand ,
grow , edg ,
senior , appoint , Etihad ,
Officer , Chief ,
appoint , Dubai ,
appoint ,
Chief , Execut , CEO , grow ,
appoint ,
resign ,
Execut , Chief , Non ,

Show words

period bank effect connect compon busi billion appoint execut director current vice presid parti secretari gener automot group hold compani limit control sharehold brillianc china entitl annual remuner announc wholli own subsidiari automobil

Applications

Text Summarization (Sentences X Words)

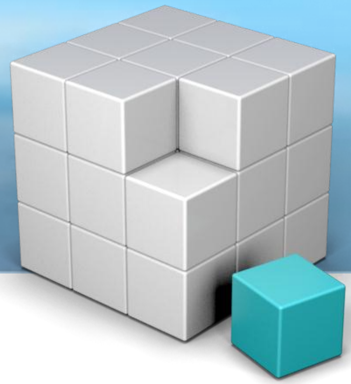


Micro Pseudo Coverage

(File: D:\test\web\mana-12.txt)

The International AIDS Vaccine Initiative (IAVI) congratulates its founder, President and Chief Executive Officer Seth Berkley on his appointment as Chief Executive Officer of the Global Alliance for Vaccines and Immunization (GAVI). The Board of GAVI, an organization that increases access to immunization in poor countries, announced the appointment today."This is a terrific opportunity for Seth and a logical next step, and we wish him well," said IAVI Board Chairman Paul Klingenstein.Berkley founded IAVI in 1996.Since then, the organization has mobilized US$ 897 million in resources for AIDS vaccine R&D, advocated to keep the development of an AIDS vaccine on the international agenda, established and maintained a network connecting clinical research centers in Africa and India, tested nine AIDS vaccine candidates in clinical trials in 11 countries and made significant contributions toward solving the problem of how to design vaccines that elicit antibodies that neutralize a broad range of variants of HIV."It was Seth's vision and energy that enabled IAVI to prosper and develop into the mature organization that it is, and he will leave it in excellent condition," said Klingenstein."It's no longer a question of whether there will be an AIDS vaccine but when, and IAVI is in an excellent position to help advance the science to the goal of an effective vaccine accessible to all."I'm excited about the new opportunity at GAVI, but of course extremely saddened to be leaving an organization so close to my heart.It's bittersweet to leave IAVI when AIDS vaccine science is showing more promise than ever.I look forward to the time when HIV will be added to GAVI's list of available vaccines," said Berkley."It's been a tremendous honor to have worked alongside all the outstanding individuals who have taken up the mission of an AIDS vaccine.I especially want to thank IAVI's generous donors, the trial volunteers and our scientific collaborators."Seth has been an indefatigable and enormously effective champion for AIDS vaccine development," said David Cook, Chief Operating Officer at IAVI."We expect he will continue to be a strong advocate from his new position at GAVI.Berkley is expected to remain in his position as President and CEO of IAVI through June.The IAVI Board of Directors will define a transition plan."Given the momentum in AIDS vaccine development and IAVI's contributions to the current excitement, a transition presents the opportunity to invigorate the organization with another outstanding leader," said Klingenstein."The Board looks forward to rising to that challenge. About IAVI The International AIDS Vaccine Initiative (IAVI) is a global not-for-profit organization whose mission is to ensure the development of safe, effective, accessible, preventive HIV vaccines for use throughout the world.Founded in 1996 and operational in 25 countries, IAVI and its network of collaborators research and develop vaccine candidates.IAVI's financial and in-kind supporters include the Bill & Melinda Gates Foundation, the Foundation for the

Future work



- Use Ontology aligned structuring techniques

- Use higher abstractive approximation of context coverage (Union of maximal rectangles) for obtaining a higher semantic level in the structure at a relatively low cost → Hyper rectangles

- Define optimization factors for identifying optimal update mechanisms(incremental vs. non-incremental)

Thank You
Q & A

Incremental Pseudo Rectangular Organization of Information Relative to a Domain