



Learning Contextual Behavior of Text Data

Dr. Coskun Bayrak

Hemant. Joshi

University of Arkansas at Little Rock

Presented at ICMLA 2005 conference

Contents

- Information Retrieval Techniques
- Weighting Schemes
- Contextual correlation weighting
 - Terminology
 - Information Correlation
 - Correlation Representation
- Results
- Conclusion

Information Retrieval (IR) Techniques

- Preprocessing data
 - Prune common words / stop words
 - Extract Unique words
 - Stemming
- Vector Space Model (VSM)
 - Matrix-representation
 - Rows vs. Columns
- Latent Semantic Indexing (LSI)
- Matrix decompositions
 - *Singular Value Decomposition (SVD)*
 - *Semi Discrete Decomposition (SDD)*

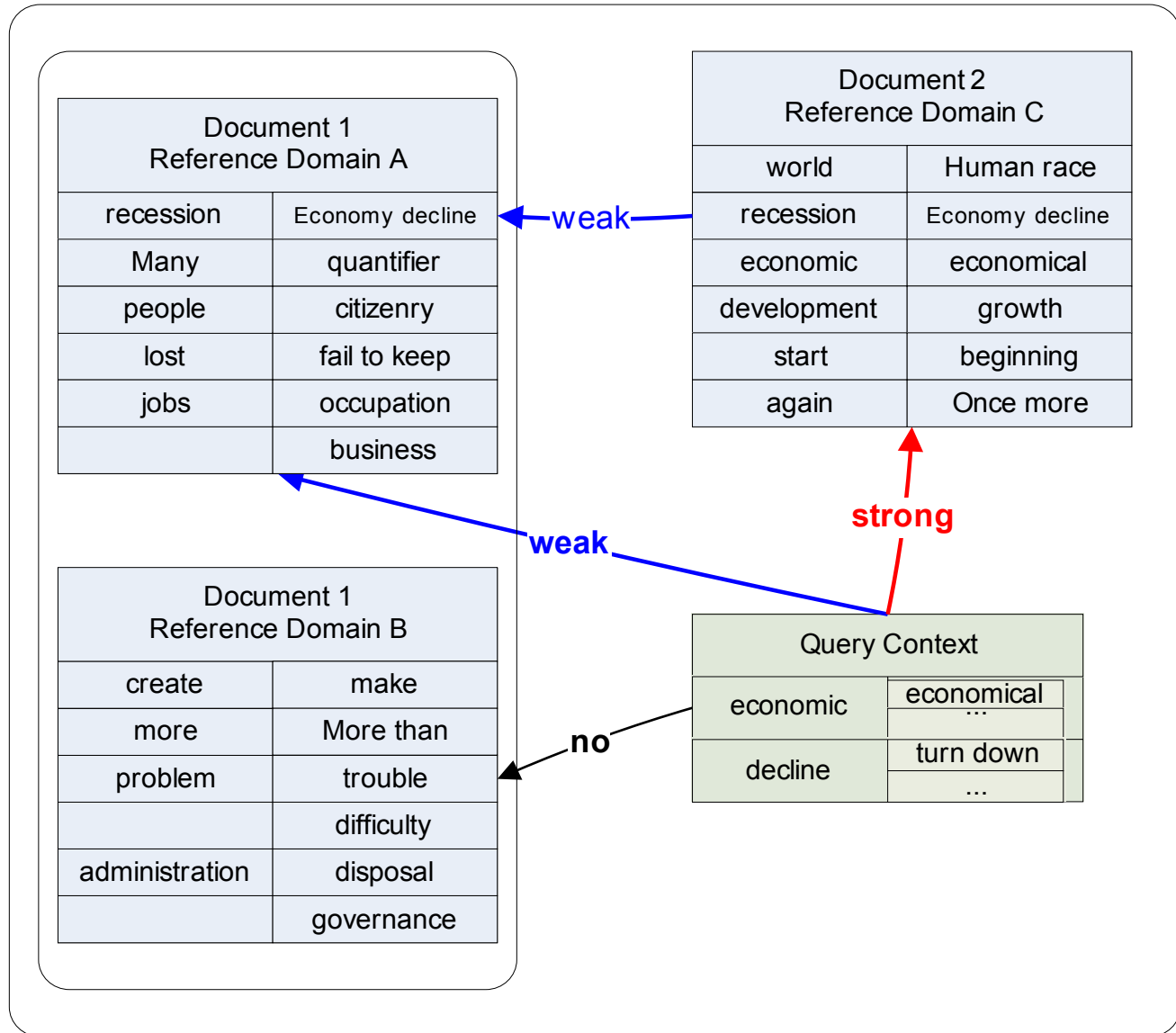
Weighting Schemes

- Term Frequency (TF)
 - Which words occur more often than others?
- Inverse Document Frequency
 - Reverse look up
- Several different weighting schemes
 - Different term weighting techniques
 - Query weighting
 - Feedback based weighting
- Contextual weight ?
 - What you want depends on what you are looking for.
 - Correlation to the query

Contextual correlation weighting

- Information Domain(s) and sub-domains
- Reference domains
- Correlation between reference domains
- Types of correlation
 - Weak correlation
 - Strong correlation
 - No Correlation
- Related Domains
 - Semantic Correlation is defined among related domains

Example : Correlation



Information Correlation & its representation

$$\Delta_i = \left[\frac{N}{N+1} \right] * \left(\frac{1+T_s}{1+N} \right)$$

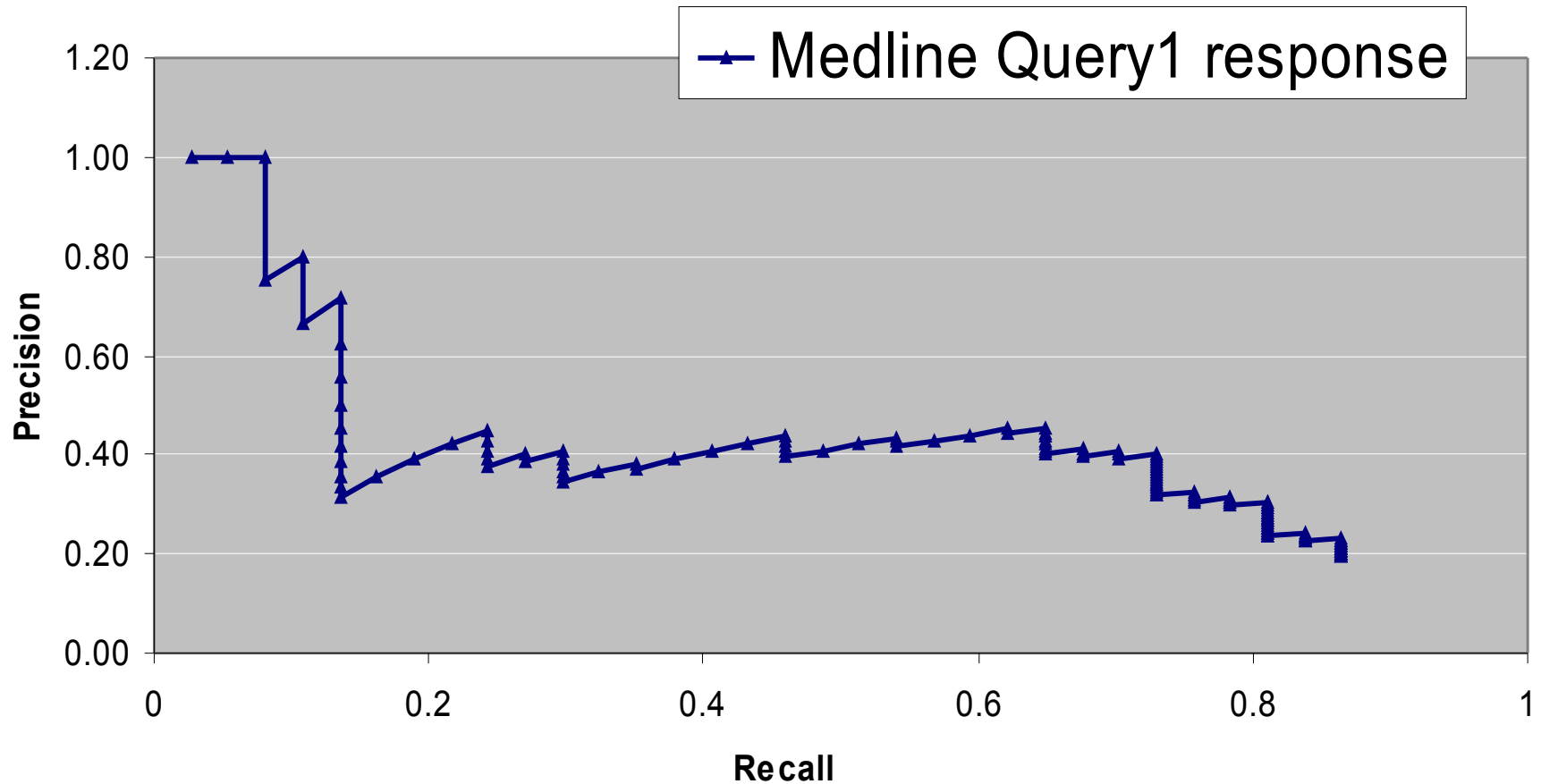
- Given reference domain i
- Δ_i is the correlation coefficient for ref. domain
- N is the number of total references to ref. domain
- T_s is the number of strong references to I
 $\therefore N = T_s + T_w$

where T_w is number of weak references to i

- $\Delta_i = 1 \quad \Rightarrow$ "Strong" correlation
- $\Delta_i = 0 \quad \Rightarrow$ "No" correlation
- $0 < \Delta_i < 1 \quad \Rightarrow$ "Weak" correlation

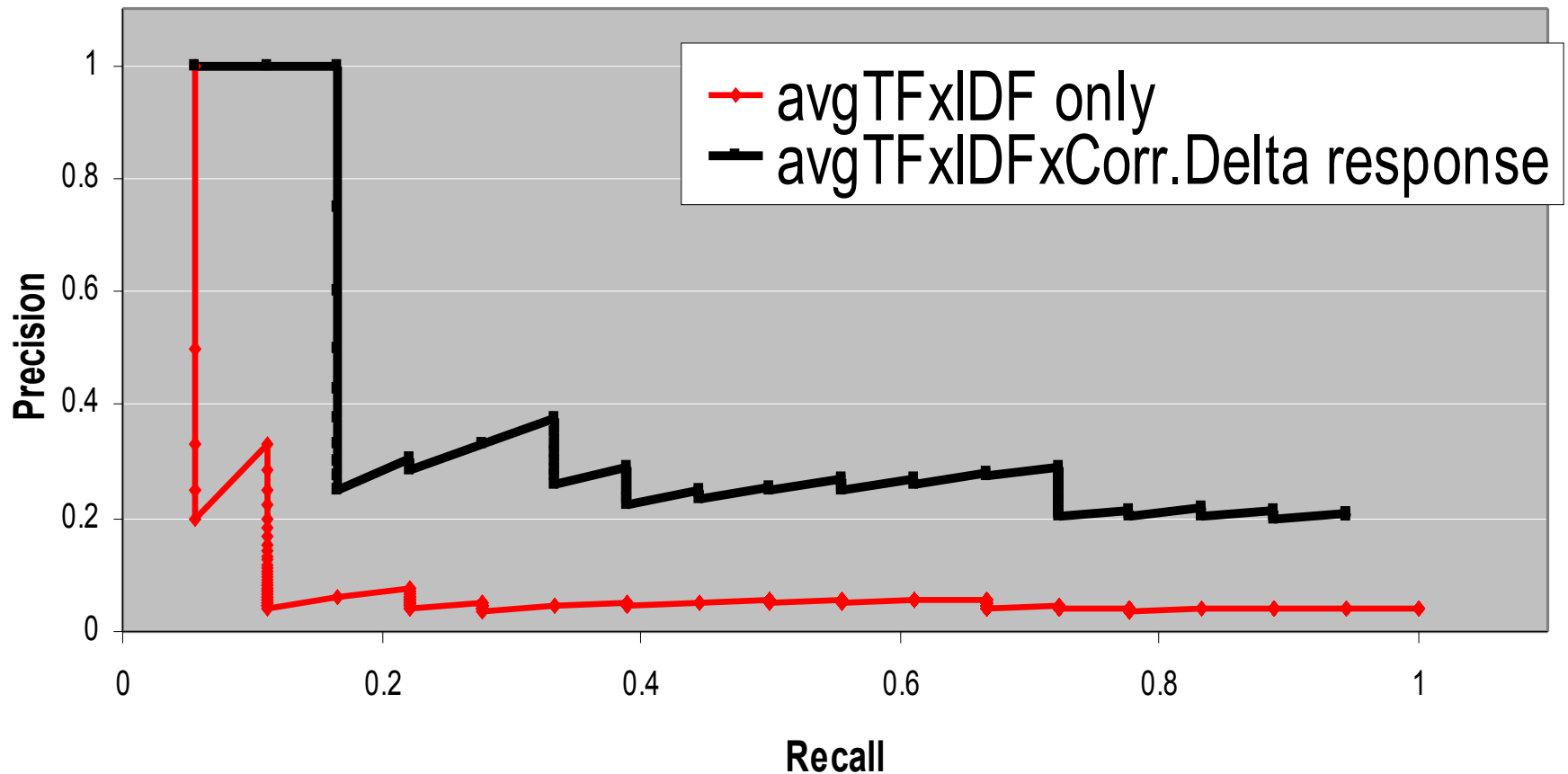
Precision Recall – MEDLINE Query 1

Precision Recall Response



Time Dataset Query 46

Precision Recall Comparison



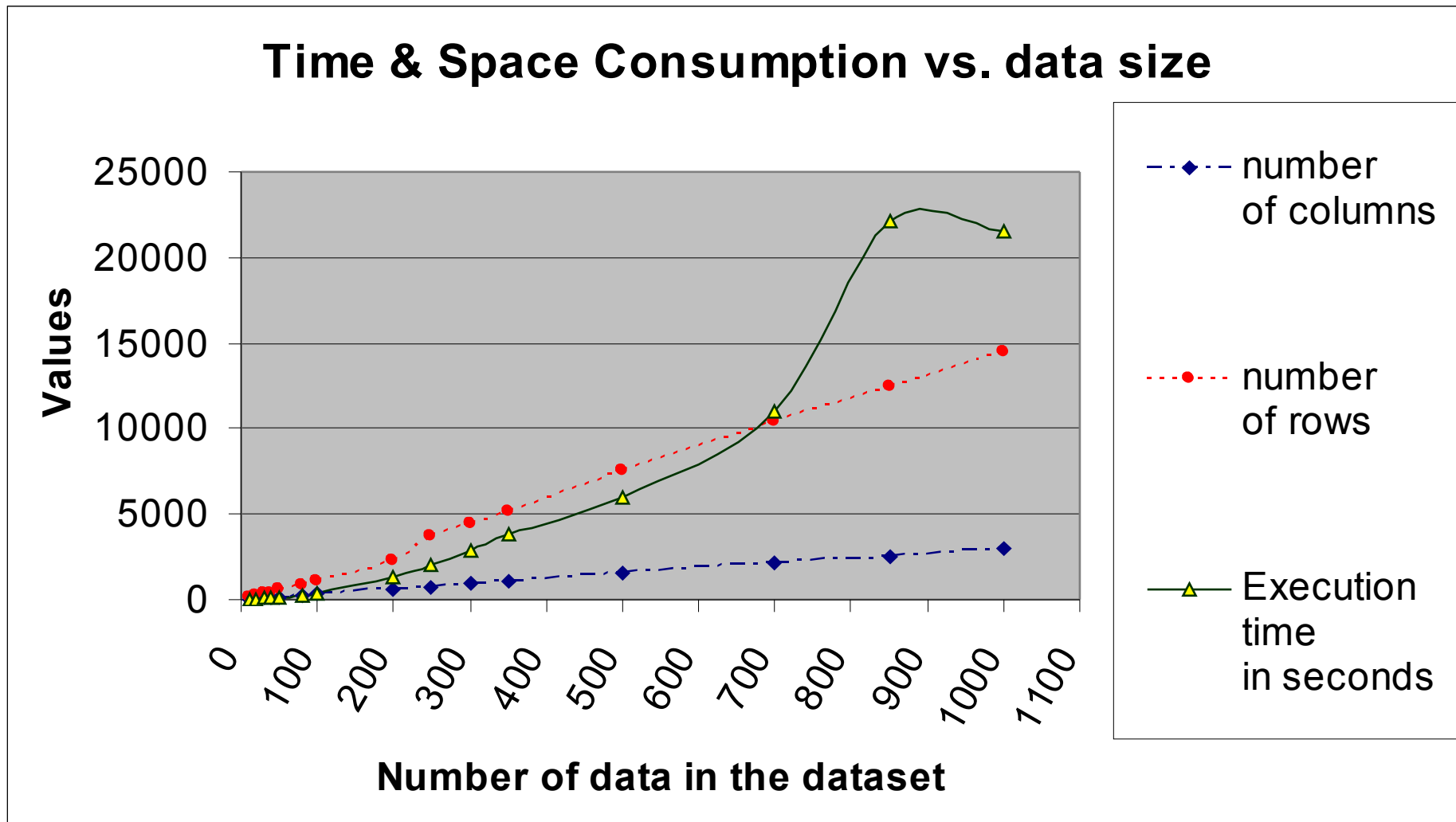
Conclusion & Future work

- Information is valid in the given context
- Contextual Weighting is important
- Comparable IR performance
- Future Work
 - Large scale testing
 - Disambiguation for more accurate weighting
 - Code Optimizations
 - Use with LSI and SVD/SDD

Thank you

Results: Time-space complexity

The results were obtained with Windows XP Pro/Linux Mandrake 10.0, 1GB RAM, Intel 2.4 GHz. Dual Processor, Java



Comparison

Precision Recall Comparison

