Toward Entity Retrieval over Structured and Text Data

Mayssam Sayyadian, Azadeh Shakery,
AnHai Doan, ChengXiang Zhai

Department of Computer Science

University of Illinois, Urbana-Champaign

Presentation at ACM SIGIR 2004 Workshop on Information Retrieval and Databases, July 29, 2004

Motivation

- Management of textual data and structured data is currently separated
- A user is often interested in finding information from both databases and text collections. E.g.,
  - **Course information may be stored in a database; course web sites are mostly in text**
  - **Product information may be stored in a database; product reviews are in text**
- How do we find information from databases and text collections in an integrative way?

Entity Retrieval (ER) over Structured and Text Data

- Problem Definition
  - Given collections of structured and text data
  - Given some known information about a real-world entity
  - Find more information about the entity
- Example
  - Data= DBLP (bib. Database) + Web (text)
  - Entity = researcher
  - Known information = "name of researcher" and/or a paper published by the researcher
  - Goal = find all papers in DBLP and all web pages mentioning this researcher

Entity Retrieval vs. Traditional Retrieval

- ER vs. Database Search
  - **ER requires semantic-level matching**
  - **DB search matches information at the syntactic-level**
- ER vs. Text Search
  - **ER represents a special category of information need, which is more objectively defined**
- What's new about ER?

Challenges in ER

- Requires semantic-level matching
  - **Both DB search and text search generally match at the syntactic level**
  - **E.g., name= “John Smith” would return all records match the name in DB search**
  - **E.g., query=“John Smith” would return documents match one or both words**
  - **But “John Smith” could refer to multiple real-world entities**
- Same name for different entities
- A unique entity name may appear in different syntactic forms in a DB and text collection.
  - **E.g., “John Smith” -> “J. Smith”**

Definition of a Simplified ER Problem

Query $Q=(q, R, C, T)$

q=*Text query*

$R=\{r_1, r_2, \ldots, r_m\}$ *examples of rel docs*

$r_i \in D$

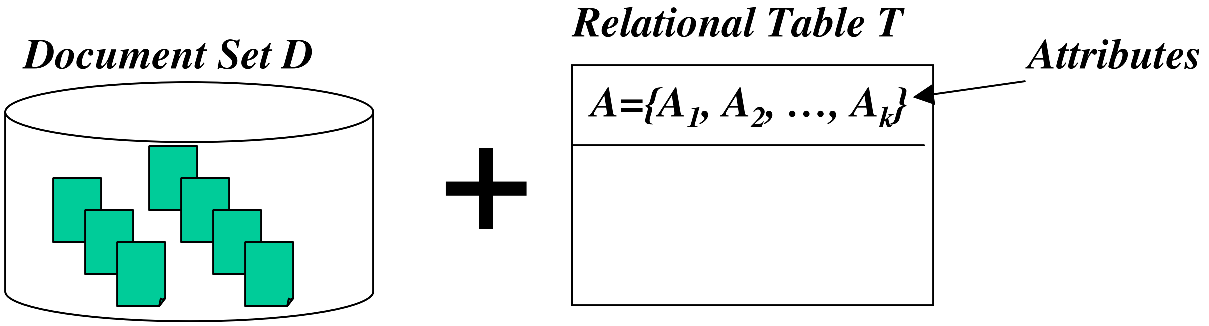
$C=\{c_1=v_1, c_2=v_2, \ldots, c_n=v_n\}$ *constraints*

$c_i \in A$

$T=\{t_1, t_2, \ldots, t_l\}$ *target attributes*

$t_i \in A$

Data



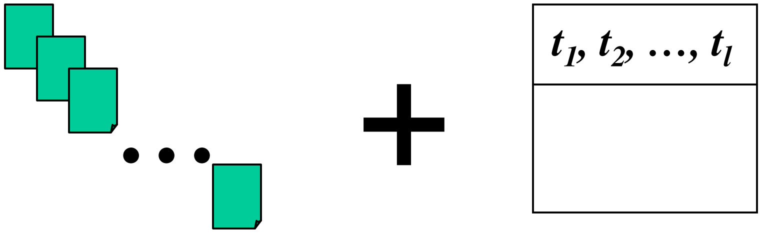
Document Set D

Relational Table T

Attributes

$A=\{A_1, A_2, \ldots, A_k\}$

Results



$t_1, t_2, \ldots, t_l$

Finding all Information about “John Smith”

Query *Q=(q, R, C, T)*

q=“John Smith”

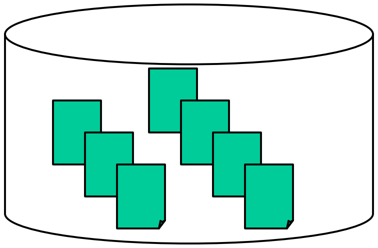
R: Home page of “John Smith”

C: {author=“John Smith”, paper.conferenc=SIGIR}

T: {paper.title, paper.conference}

Data

The Web



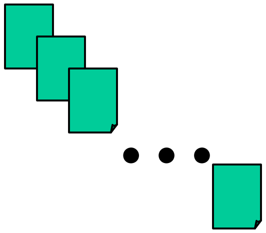
+

DBLP bib. database

Author, title, conf, date…

“John Smith” is highly ambiguous!

Results



+

Titl	*conf*

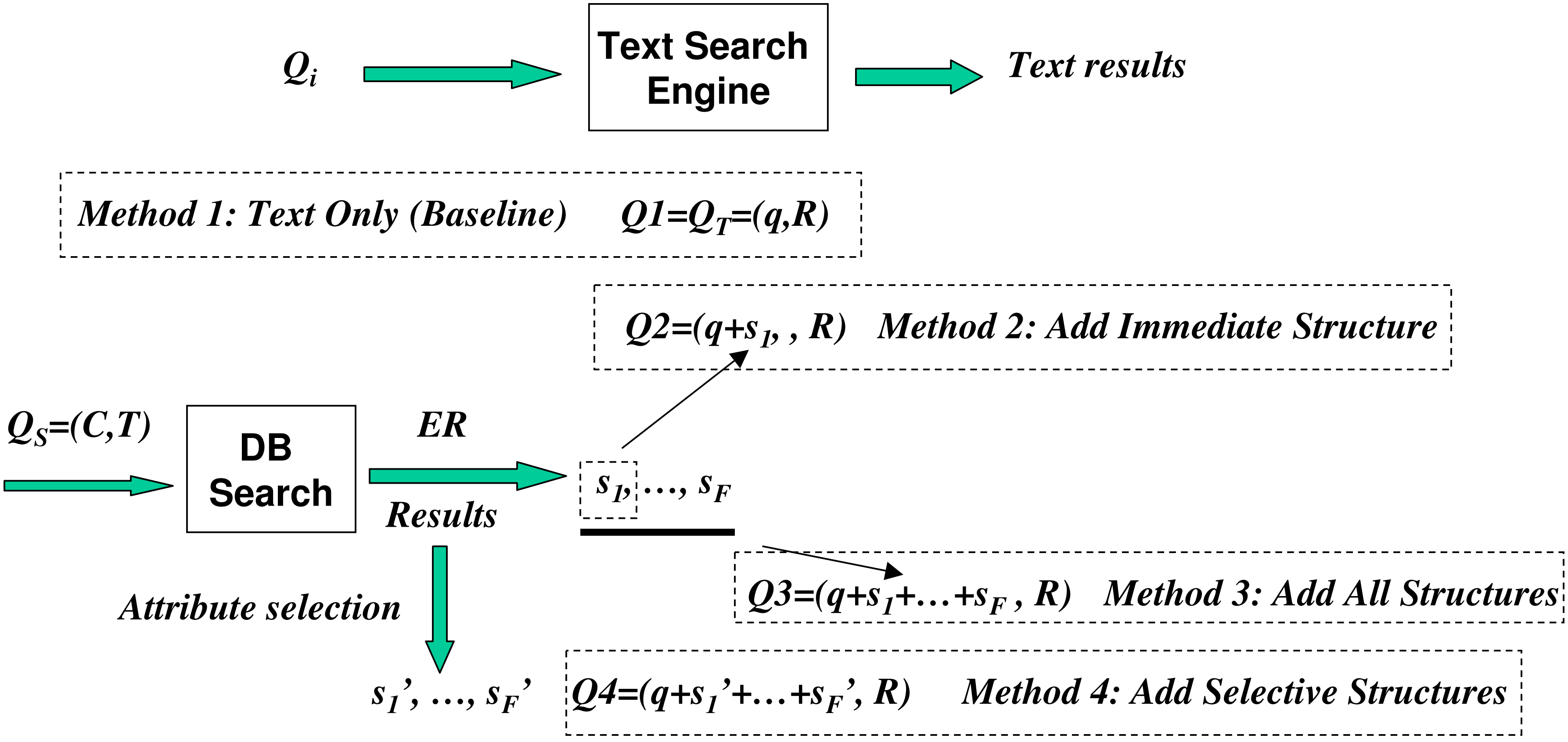
ER Strategies

- Separate ER on DB and on text
  - **Q=(q,R,C,T)**
    - Use Q1=(q,R) to search the text collection
    - Use Q2=(C,T) to search the DB
  - **The main challenge is entity disambiguation**
- Integrative ER on DB + Text
  - **Q=(q,R,C,T): use Q to search both the text collection and DB**
  - **Relevant information in DB can help improve search over text**
  - **Relevant information in text can help improve search over DB**

Hypothesis tested in this work

Exploit Structured Information to Improve ER on Text

Given an ER query Q=(q,R,C,T)
Assume that we have a basic text search engine
We may exploit structured information to construct a different text query Q_i



Q_i

Text Search Engine

Text results

Method 1: Text Only (Baseline) *$Q1=Q_T=(q,R)$*

$Q2=(q+s_1, , R)$ *Method 2: Add Immediate Structure*

$Q_S=(C,T)$

DB Search

ER

Results

$s_1, \ldots, s_F$

$Q3=(q+s_1+\ldots+s_F , R)$ *Method 3: Add All Structures*

Attribute selection

$s_1', \ldots, s_F'$

$Q4=(q+s_1'+\ldots+s_F', R)$ *Method 4: Add Selective Structures*

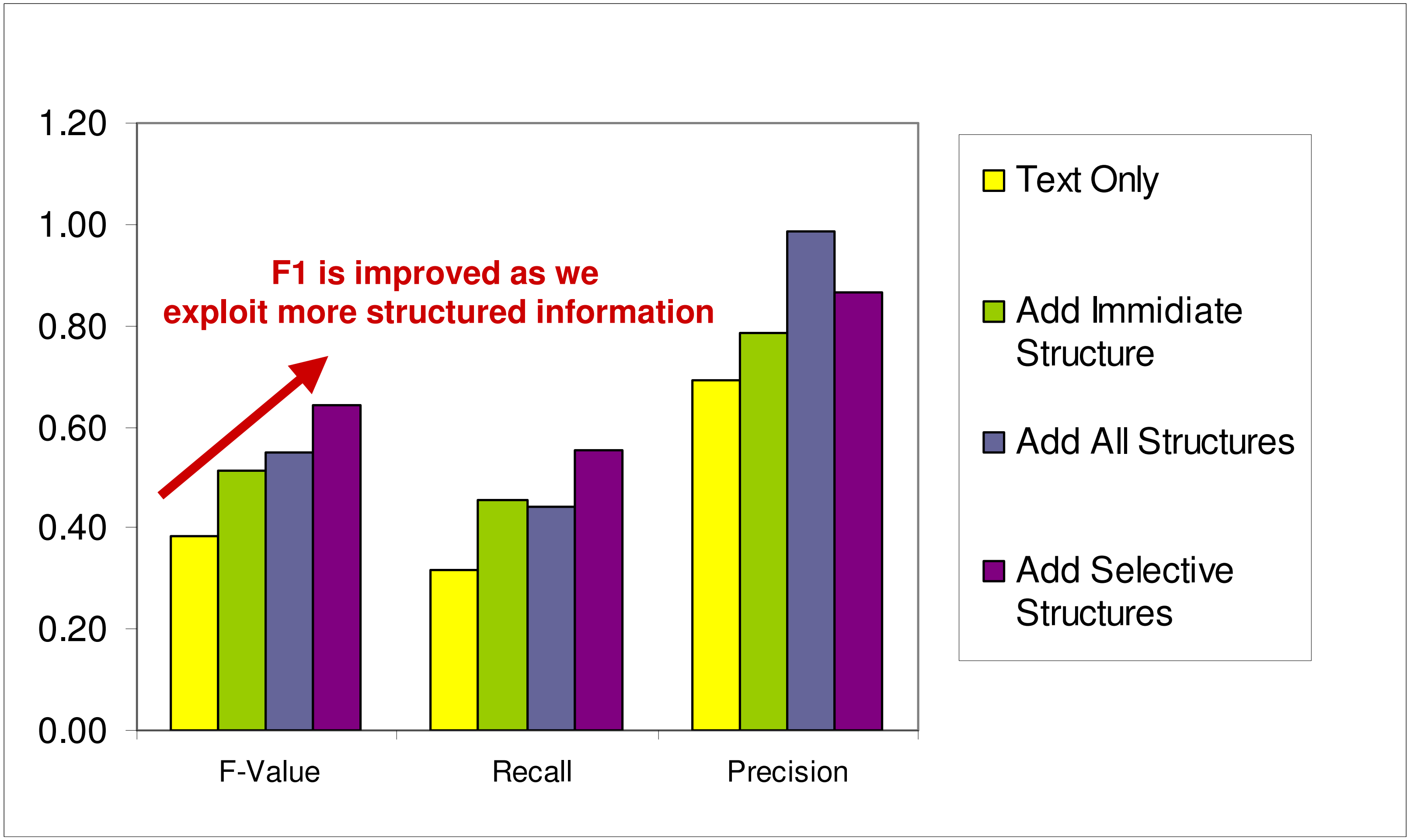
Attribute Selection Method

- Assumption: An attribute is more useful if it occurs more frequently in the top text documents (returned by the baseline TextOnly method)
- Attribute Selection Procedure
  - **Use the top 25% of the docs returned by TextOnly as the reference doc set**
  - **Score each attribute by the average frequency of all the attribute values of the attribute in the reference doc set**
  - **Select the attribute with the highest score to expand the query**

Experiments

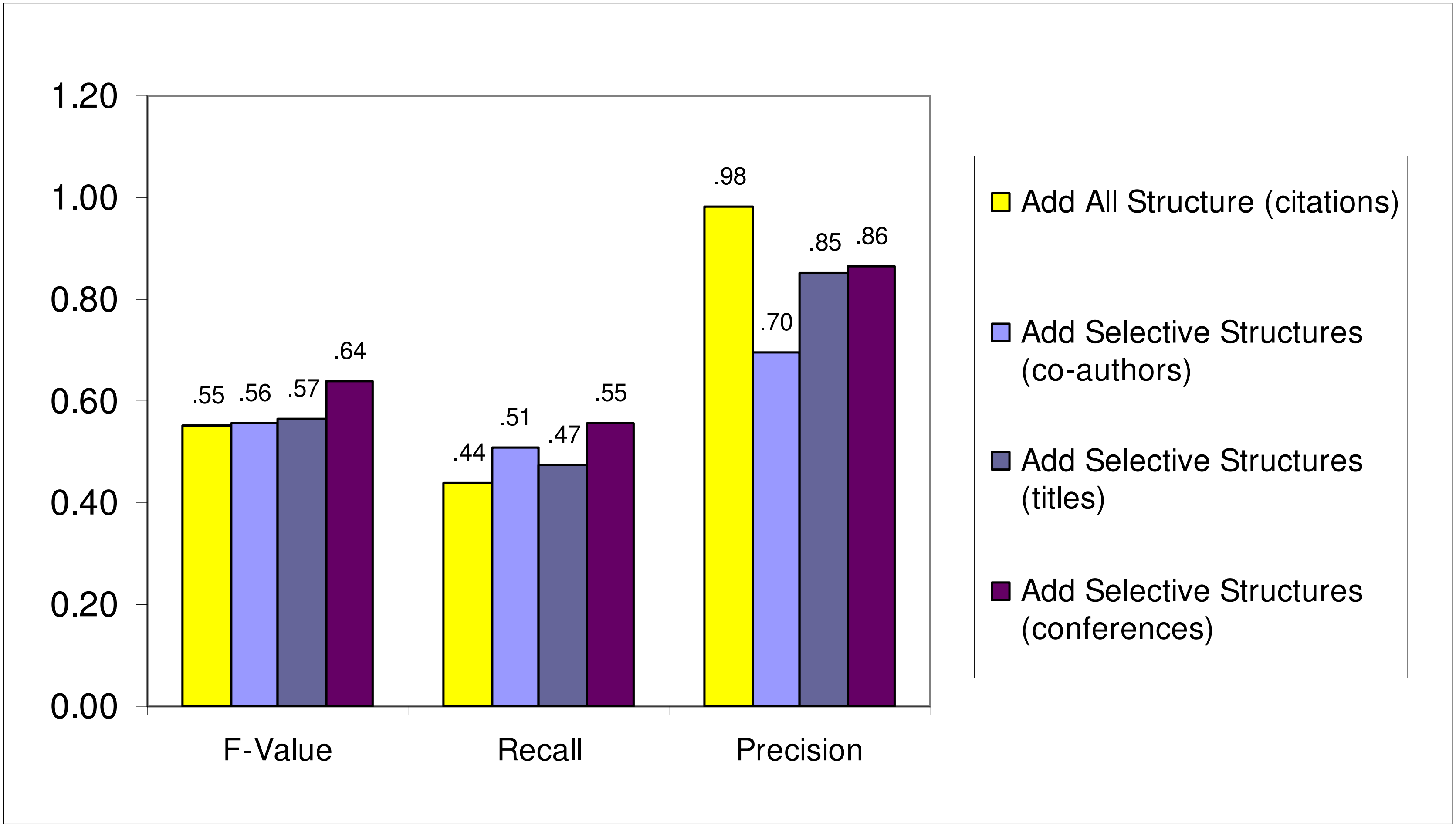
- ER queries: 11 researchers, Q=name (no relevant text doc examples)
- DB = DBLP (www.informatik.uni-trier.de/ley/db) , >460,000 articles
- Text collection = top 100 web pages returned by Google using the names of the 11 researchers
- Measures:
  - **Precision: percent of pages retrieved that are relevant**
  - **Recall: percent of relevant pages that are retrieved**
  - **F1: a combination of precision and recall**
- Retrieval method
  - **Vector space model with BM25 TF**
  - **Scores normalized by the score of the top-ranked document**
  - **A score threshold is used to retrieve a subset of the top 100 pages returned by Google (set to a constant all the time)**
  - **Implemented in Lemur**
- ER on DB: the DBLP search engine on the Web with manual selection of relevant tuples

Effect of Exploiting Structured Information



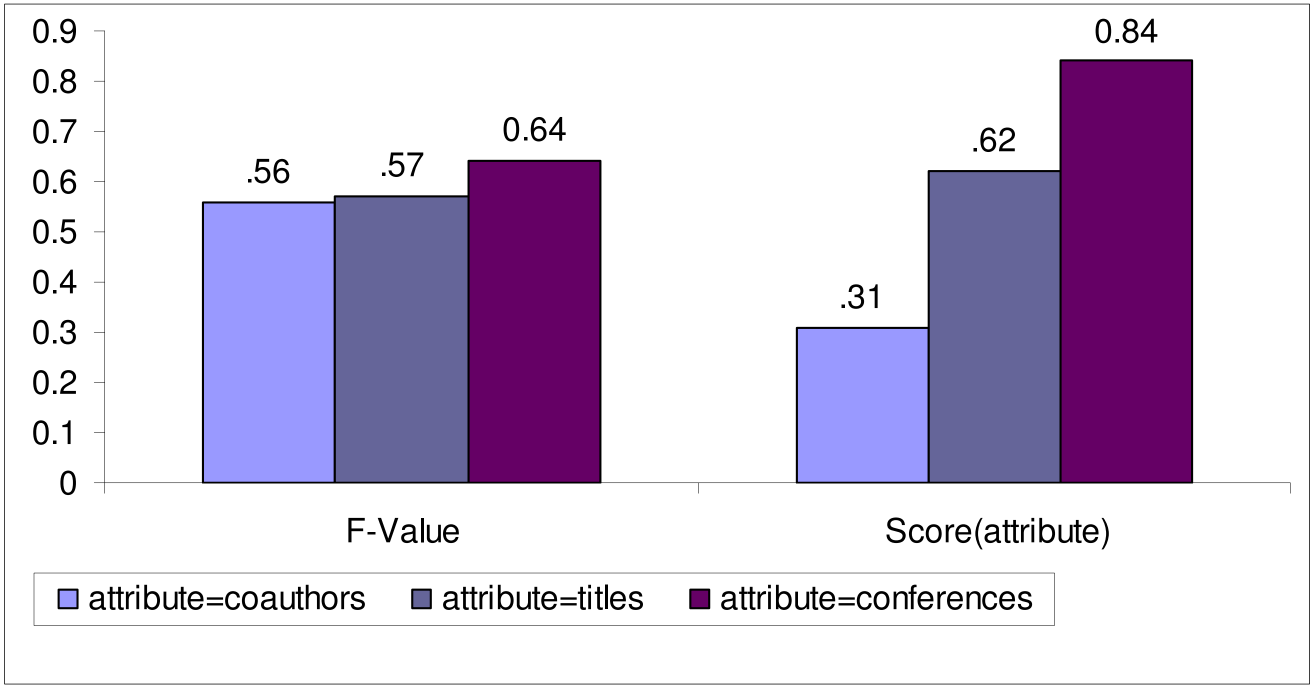
F1 is improved as we exploit more structured information

Effect of Attribute Selection



Conference is a better attribute than co-authors or titles

Automatic Attribute Selection



The attribute score based on value frequency predicts the usefulness of an attribute well

Conclusions

- We address the problem of finding information from databases and text collections in an integrative way
- We introduced the entity retrieval problem and proposed several methods to exploit structured information to improve ER on text
- With some preliminary experiment results, we show that exploiting relevant structured information can improve ER performance on text

Many Further Research Questions

- What is an appropriate query language for ER?
- What is an appropriate formal retrieval framework for ER?
- What are the best strategies and methods for ER?
- …

Thank You!