



Fuzzy Multi-Dimensional Search in the Wayfinder File System

Christopher Peery, Wei Wang, Amélie Marian, Thu D. Nguyen
Computer Science Department, Rutgers University



Motivations

- Current desktop search tools use:
 - Keyword search for **ranking** results
 - Other dimensions such as metadata and structure for **filtering** results
- This approach is **insufficient** for personal file search

Example Query:

[FileType = *.doc" and Content = "paper draft" and CreateDate = "11/10/07" and Path = "/pubs/submit"]

Current tools **will not** return relevant files created **around** 11/10/07, **similar** to type .doc, or **stored near** "/pubs/submit"

- Users may remember **imprecise information** about target files
 - Limited use for filtering
 - Highly useful to guide **fuzzy search**

Contributions

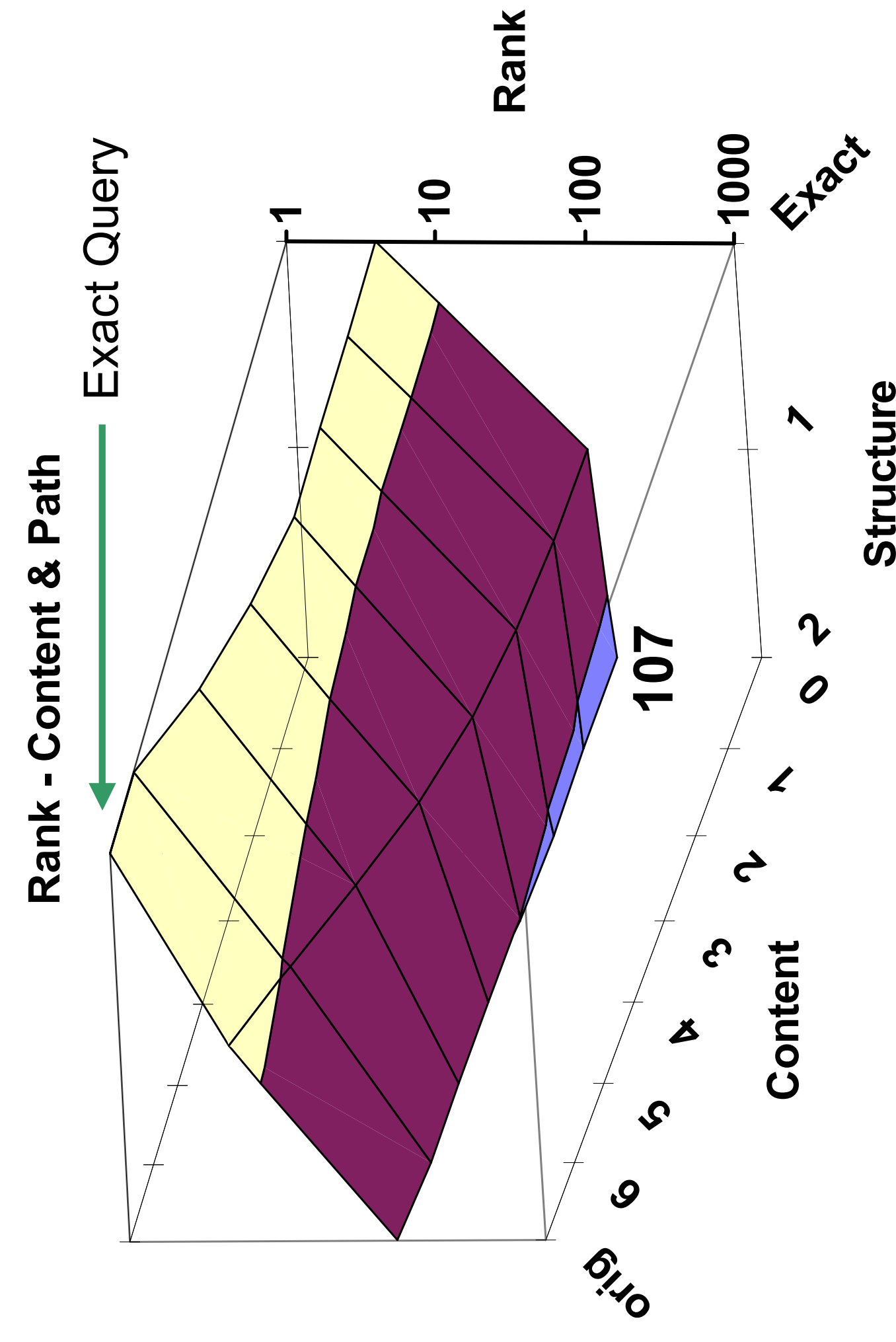
Improve the quality of file system search by:

- Allowing **approximate matches** to query conditions
- Ranking results based on their relevance across **multiple dimensions**

Unified Scoring Framework [EDBT08][SEARCH08]:

- **Relaxation approaches** defined for conditions in multiple dimension: content, metadata, and structure
- **IDF-based** scoring to support ranking of approximate matches
 - Score decreases as number of matching files increases
- Individual dimension scores are aggregated into a **single relevance score**

Behavior of Inaccurate Queries – 2D



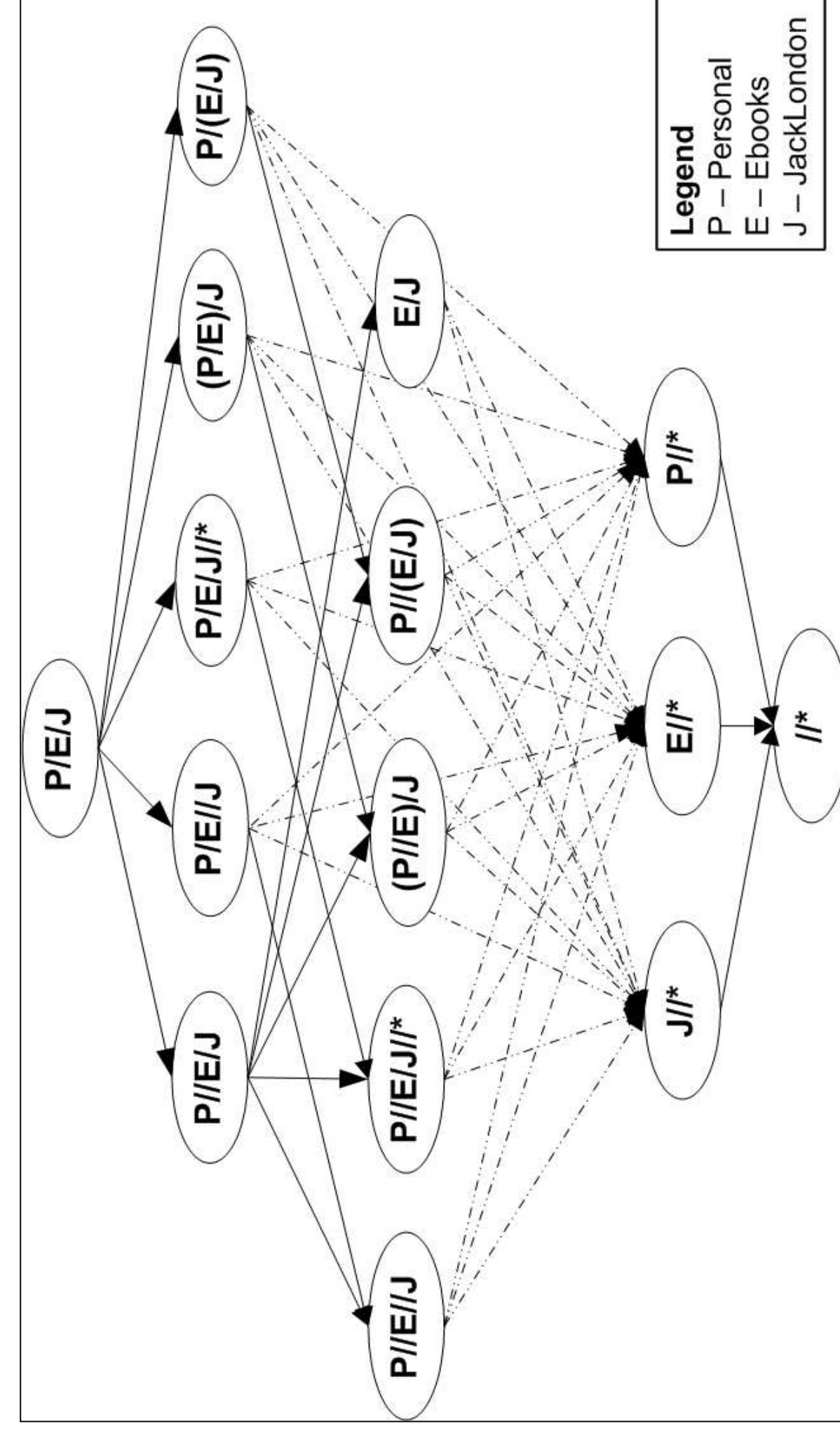
Structure Scoring

- Structure relaxation based on **XML structure query relaxations**

- 1) **Edge Generalization:** /a/b/c ⇔ /a/b/c
- 2) **Path Extension:** /a/b ⇔ /a/b/*
- 3) **Node Deletion:** /a/b/c ⇔ /a/c
- 4) **Node Inversion:** /a/b/c ⇔ /a/(b/c) [a/b/c or /a/c/b]

- All relaxation compositions represented in a DAG
 - Root represents exact match, leaf node represents relaxed form that matches all directories
 - Each combination represented by node and has associated **IDF score**
 - DAG is **query dependent** and computed at query time
- Indices/Algorithms developed to optimize the creation and evaluation of DAGs
 - **Exponential** number of combinations

Structure Index



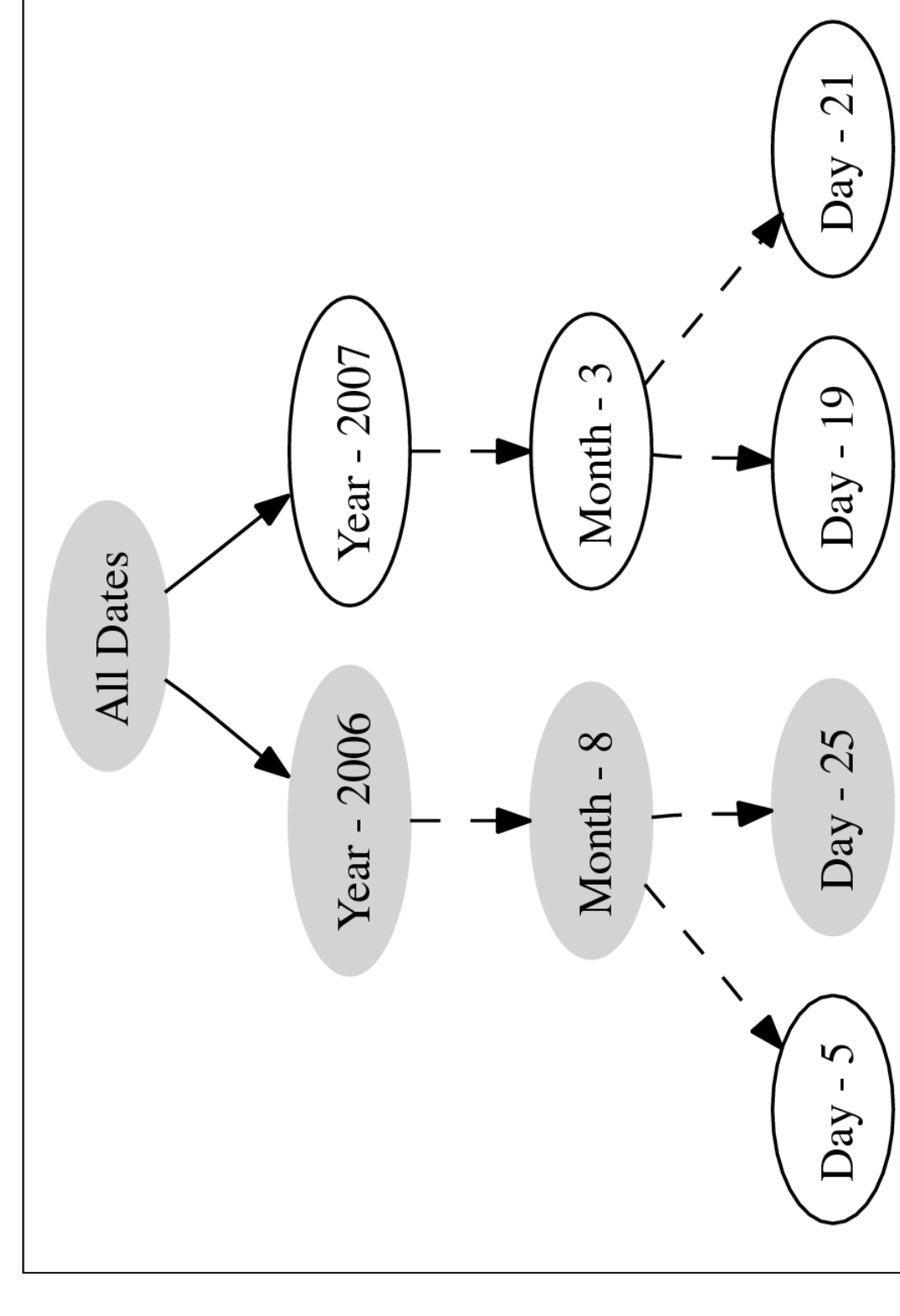
Query Path = /Personal/Ebooks/JackLondon

Metadata Scoring

- Approximations are generalizations of query values
 - **Date:** Exact Time → Day → Month → Year
 - **File Type:** PDF → Documents → All Files
 - IDF score associated with each approximation
- Relaxations/Index are represented as a DAG
 - Leaf to root node represents exact match to most general approximation
 - **IDF scores** decrease from leaves to root

Metadata Index

Example Date Relaxation DAG:



Date Query = August 25, 2006

Content Scoring and Indexing

- Keyword conditions scored using standard **TFxIDF** over inverted indices

Aggregation

- IDF-based dimension scores measure similar information
- Vector projection used to aggregate scores from multiple dimensions

Impact of Flexible Multi-Dimensional Search

Query	Query Conditions (Content, Type, Path)	Rank	Comment
Q1	C	49	Base Query
Q2	C .txt	2	Correct Value (all Dim.)
Q3	C .txt /p/e/n/j	6	Correct Value
Q4	C .doc	45	Incorrect Value
Q5	C Docs.	21	Relaxed Value
Q6	C /p/e/n/j	3	Correct Path
Q7	C /j/e	3	Incorrect Path
Q8	C Docs. /p/e/n	15	Relaxed (all Dim.)
Q9	C Docs. /j/e	2	Incorrect & Relaxed
Q10	C .pdf /j/e	2	Incorrect Value (all Dim.)

Evaluation Data Set

- **Real User** data set (approx. 27,000 files)
- Queries of varying dimensions and relaxations; built around specific target files

Citations

- [EDBT08] C. Peery, W. Wang, A. Marian, and T. D. Nguyen. **Multi-Dimensional Search for Personal Information Management Systems**.
- [SEARCH08] W. Wang, C. Peery, A. Marian, and T. D. Nguyen. **Efficient Multi-Dimensional Query Processing in Personal Information Management Systems**. Technical Report DCS-TR-627, Dept. of Computer Science, Rutgers University, 2008.