A Pointer-Based File Management System to Reduce Redundancy and Storage Overhead

Braden D. Licastro
Department of Computer Science, Allegheny College

Motivation

- There was approximately 2.8 billion terabytes of data in the world as of January 2012.
- Approximately 5% of the world's data is redundant.
- Collectively, businesses spend $1.8 trillion annually to store data.
- By removing even 1% of the redundant data, that is an annual savings of $18 billion dollars.

What is a file system?

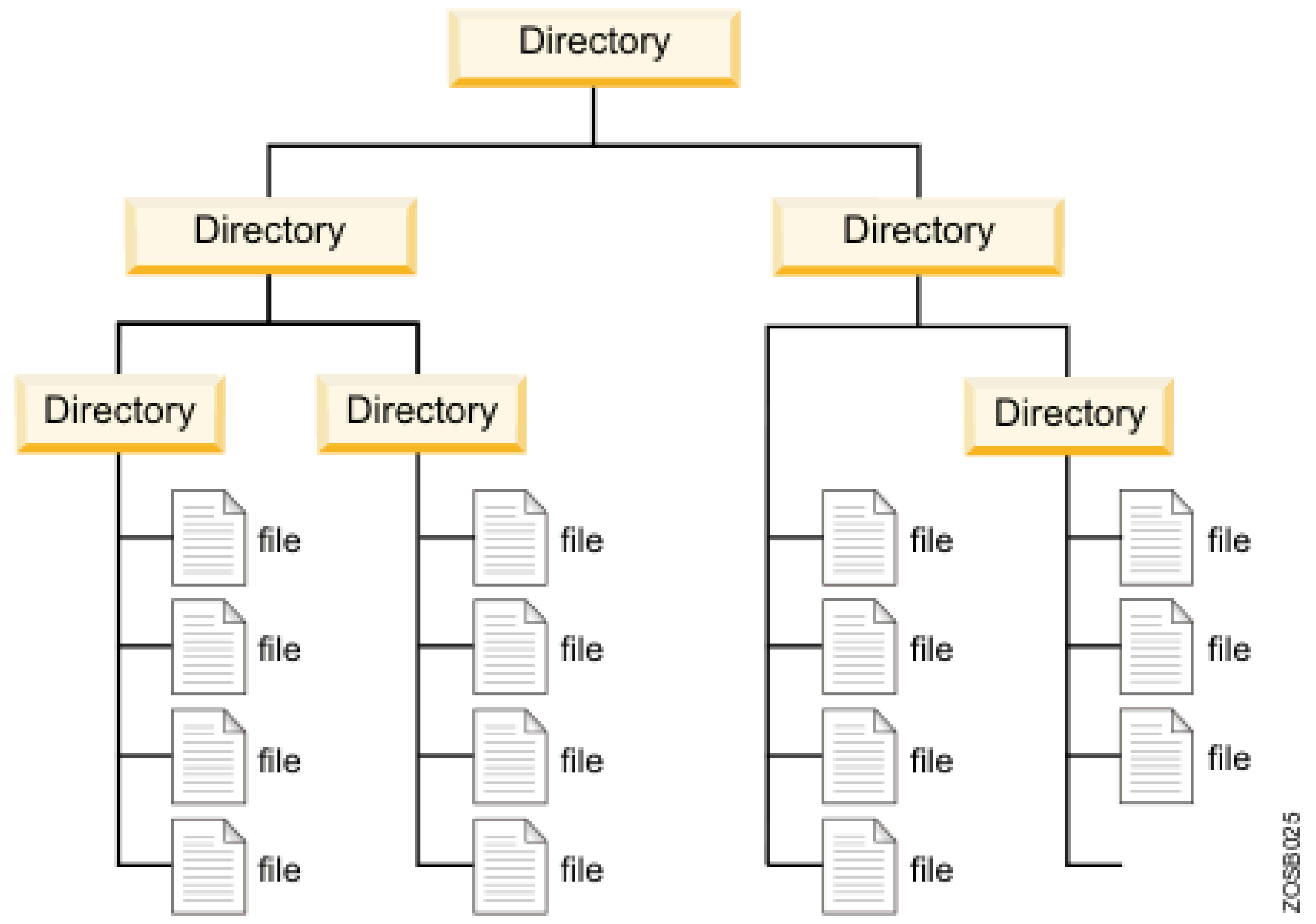


Figure: Example file system tree structure.

File systems are a type of data store that can be used to store, retrieve, and update a set of files on a hard disk.

Types of file systems

- ntfs
- exfAT
- fat
- fat32
- ext3

- Why so many? Each file system is designed and tailored for a specific need.
- All-inclusive file system would not be user friendly and would more than likely become bloated and unreliable.

Related Work

A Fast Filtering Scheme for large Database Cleansing [1]

- Database focused algorithms
- Data redundency reduction
- Matches data, not entire files
- Completely removes duplicates from database

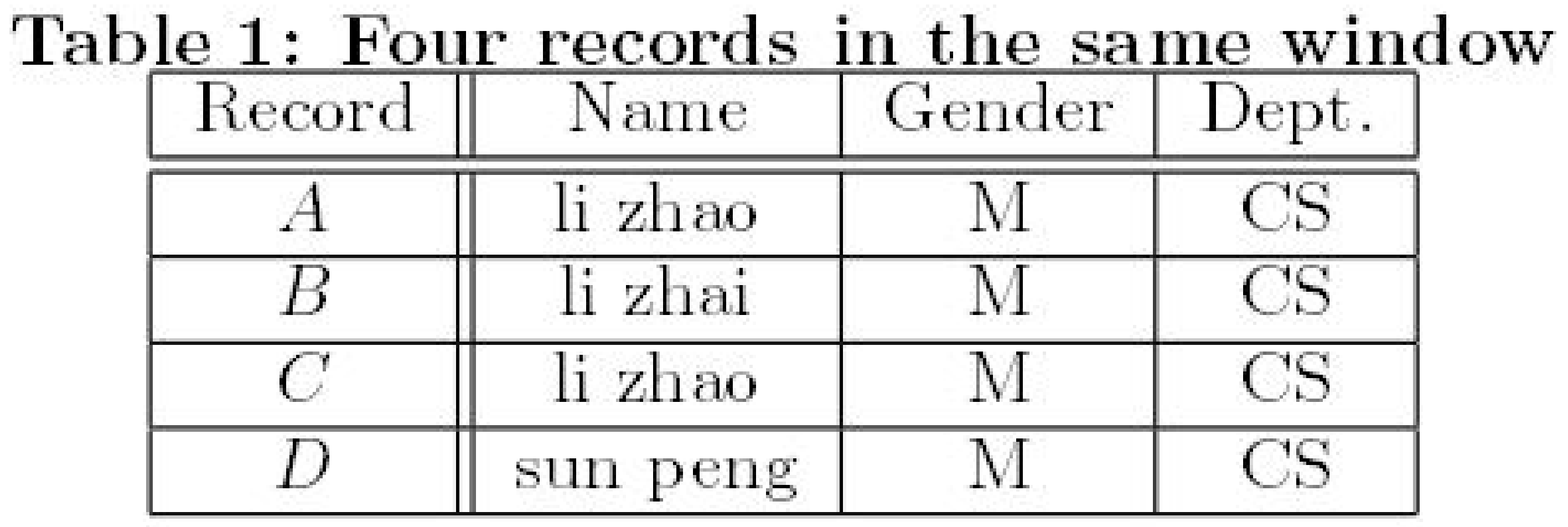


Table 1: Four records in the same window

Record	Name	Gender	Dept.
A	li zhao	M	CS
B	li zhai	M	CS
C	li zhao	M	CS
D	sun peng	M	CS

Related Work

A Data De-duplication Access Framework for Solid State Drives [2]

- Algorithms tuned for solid state drives running on computing cluster
- Finds candidate duplciates from calculated scores
- Minimal read and write calls needed to operate

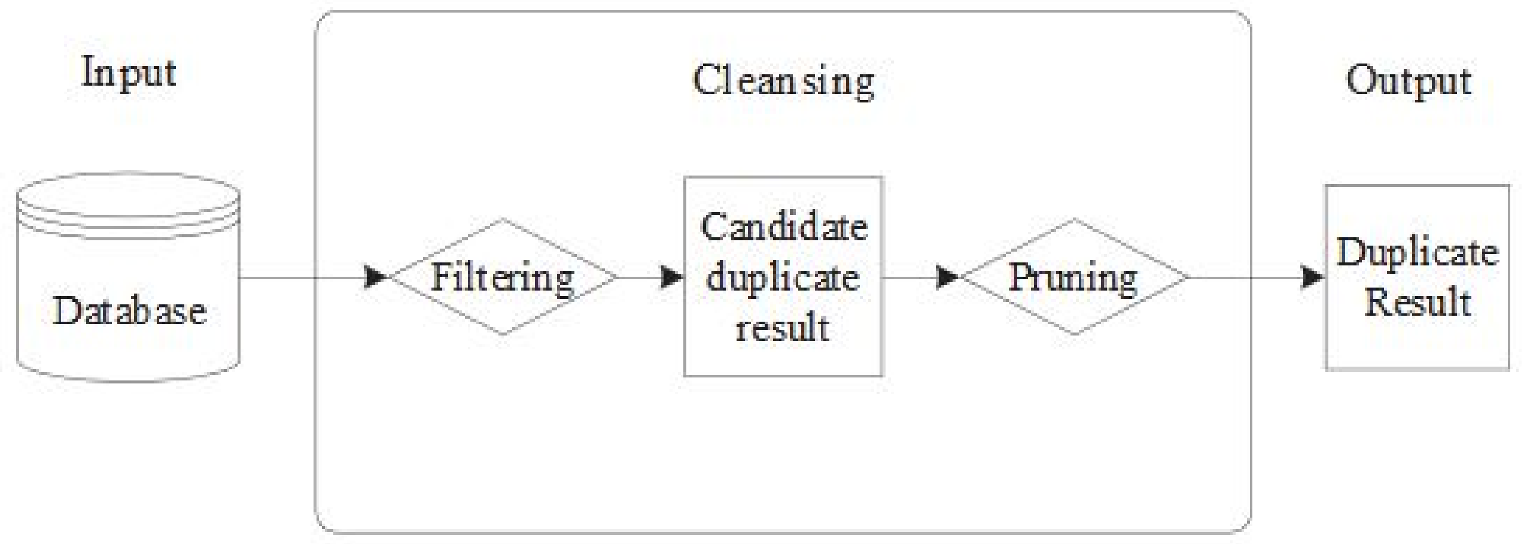


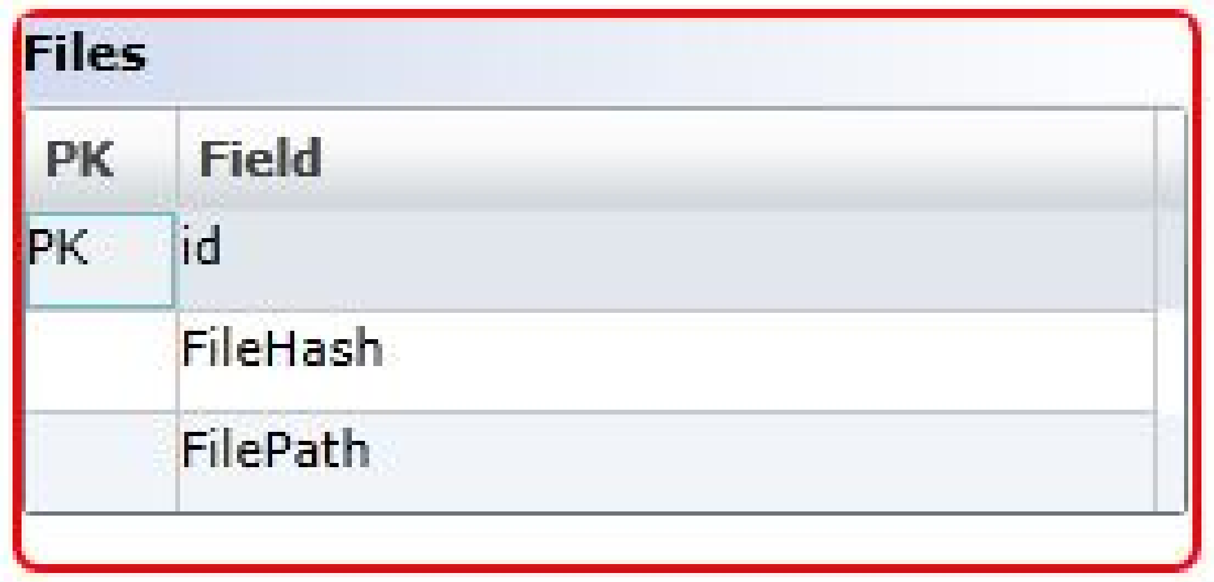
Figure: Authors process for finding and processing duplicates.

Challenges

- Initial setup time costs when hashing user files.
- Keep performance costs to a minimum when interacting with system.
- Implementing a web based environment for file system interaction.
- Verify that pointers all function as intended and file modifications drop pointer accordingly.

Implementation

- Create or modify existing file system
- Hash files usign MD5 hashing algorithm to allow checking for matches.
- Remove duplicate files and replace duplicate with pointer referring to the parent copy
- Overwrite opened pointer with modified file to prevent overwriting a needed original.
- Restrict file database to non-system files.



Files

PK	Field
PK	id
	FileHash
	FilePath

Method of Evaluation

- Measure space saved by removing duplicates including size of the database in calculations.
- Monitor performance impact of running implemented project vs OOB file system
- View file tree structure, starting with a set structure and compare to tree after removing duplicates.
- Verify that pointers all function as intended and file modifications drop pointer accordingly.

Future Work

[1] Lesley Anderson, Dr. Jon Purdy, and Warren Viant. Variations on a fuzzy logic gesture recognition algorithm. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, ACE '04, pages 280–283, New York, NY, USA, 2004. ACM.

[2] Atsushi Shimada, Manabu Kawashima, and Rin-ichiro Taniguchi. Early recognition based on co-occurrence of gesture patterns. In *Proceedings of the 17th international conference on Neural information processing: models and applications - Volume Part II*, ICONIP'10, pages 431–438, Berlin, Heidelberg, 2010. Springer-Verlag.

http://www.fullforceapps.com

First Annual Computer Science 580 Research Symposium

licastb@allegheny.edu