GW | Office of Technology Transfer

Novel Technology Opportunity Available
Non-confidential summary

Method for Approximate Searching in Very Large Files

Inventors:

Simon Berkovich, Professor of Engineering and Applied Science,
Maryam Yammahi, ***Chen Shen,*** Department of Computer Science,
The George Washington University

Field

Computer Science,
Information Retrieval

Objective

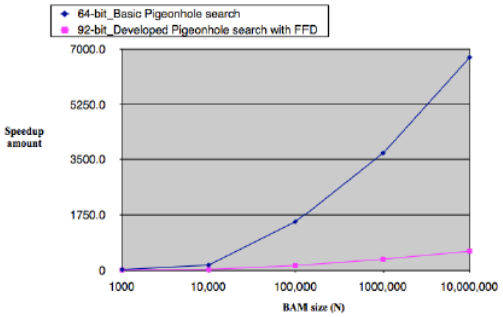
Seeking development and licensing partners

Keywords

Big Data,
Approximate searching,
Information Retrieval,
FuzzyFind dictionary,
Image recognition,
Data Structures

This invention offers an efficient means of executing “fuzzy” or approximate searches in large information systems. Approximate search for information items in extremely large data files is a very challenging problem in computer science. In big data files such as those beyond terabytes, traditional methods like sequential search substantially undermine system performance. If a sequential search is applied, it results in long search times, as time consumed is directly proportional to the size of the dataset. The novel methods created utilize the Pigeonhole Principle and some other novel techniques to speed-up the operation of approximate matching.

The basic embodiment is extremely efficient and thousands of times faster than the traditional sequential search approach. The reduction in search time versus a sequential search (“Speedup” ratio) increases as the size of the dataset growths. In our experiment (search for a 64 bit word tolerating 3 mismatches in a Nx64 matrix), as the matrix size increased from 10^3 bytes to 10^7 bytes, the “Speedup” increased by more than 5,000 as shown in the graph. This method can be applied in facial recognition systems, biometric characteristics verification, real-time speech recognition, QR code recognition, and many other current technologies in expansion.



The second embodiment incorporating the FuzzyFind Dictionary greatly increases flexibility in the search process, as this method allows significantly bigger input requests and it is tolerant to more errors in a given request. There is a tradeoff between speed and flexibility; however, this method still works 500 times more efficiently than the sequential search, and it also retains the high accuracy advantage of the first embodiment. Each embodiment holds distinctive advantages for different applications and systems sizes. Among others, QR code is a commercial application that fits perfectly in this method. According to the tests performed, this method provides us with huge efficiency and accuracy advantages for QR code searches, which hold a powerful place in the mobile marketing industry.

Any organization that searches large data files can benefit from these method. It is particularly valuable for searching the enormous troves of electronic data becoming available to businesses from the advent of “Big Data”.

Applications:

- Searching in large text archives and pictorial information
- QR code applications
- Verification of biometric characteristics
- Real-time speech recognition
- Software-defined big data storage

Advantages:

- Restructures the data for searching, thus increasing efficiency of retrieval process
- Speed; it outperforms other search techniques
- Lower cost and consumption of computational resources and energy

Patent Status: Provisional Patent Application Filed

Contact:

Gus Williamson, Licensing Associate
Tel: +1-202-994-8975
Email: gwilliamson@gwu.edu