# Offline Data Processing: Tasks and Infrastructure Support

T. Yang, UCSB 290N

#### **Table of Content**

- Offline incremental data processing: case study
- · Example of content analysis
- · Data store support



#### **Content Management**

- Organize the vast amount of pages crawled to facilitate online search.
  - Data preprocessing
  - Inverted index
  - Compression
  - Classify and partition data
- Collect additional content and ranking signals.Link, anchor text, log data
- Extract and structure content
- Duplicate detection



## **Examples of Context Extraction/Analysis**

- Identify key phrases that captures the meaning of this document.
  - For example, title, section title, highlighted words.
    HTML vs PDF
- Identify parts of a document representing the meaning of this document.
  - Many documents on the web contain a side-menu which do not contain primary material.
- Capture page content through Javascript analysis.
  - Page rendering and Javascript evaluation within a page

## **Example of Content** Analysis

- · Detect content block related to the main content of a page
  - Non-content text/link material is ignored or reduced in importance during indexing process

CNI.c	om		transform ( Mins			
SEARCH #	areas (reaso	(Assol)				
Name Page No.12	SCIENCE & S	PACE				
CA.	Anuarium plaus un	ale chark	distant and the second second			
Railtona Tauth	matchmaker	THE INCOME.				
Index.	The lands from \$100 also be	A	printed and a little			
144	Name and the Part likes of the	100				
Territor & Spine	ATLANTA Germanitation - Date	CONTRACTOR OF THE OWNER.	Save up to			
Education	and Worton, most Allow and Train		79% on			
Torest Contract	The intergraph and the state where an an an and the state of the state		P dr / w dea			
Execution Execution Property	Amain langerhole dans her Tape Taken		Last-Minute			
venu natur	Report of an Apply to class of	100 million 1				
-haven	rm.	De la constante	Cruises.			
A 1000	The behavior - 11 becavior without prog- anne finant recent than initial relies by CPU	and the second se				
25.23	begins all abases cars to spec-		Acres de la companya			
109-025	Earth Mark Spence 110	Annual New Annual Annua	Street of the			
Louis Par	The plut sald Pury Pessenities masses for the first-tase passengers.		State - Date			
Polluelle Rollin	"As an own data for the second, so allow		- and the second second			
Los Paulos	station descell, to not name there, too anamiteriate back there is the place		N.C.			
8-010-04	the organization of	100HL-MAL ALDEIS	-70			
10000	partial according to a californial frame in appartual, and colorinations accompared	Alada (Deegle) 0				
	to sum.	1468 0	Best Price			
	The appropriate and pought for the make in Adams, which meanthins can	in the second se	Guarantee			
	augusties.	Meraphilletic (Media 7042	(ar)			
	The whole sharks - manual after the main shared bits in the table allows "The Factor	preserv' - were delivered to be apportant				
	In genie hangenlaken verlaken.					
	The comparison of the special in theorem is to a write any and associate it and a Color online of the theory of their feeling stream stream, so handler of the resident topic and the stream st					
	Entry yourse about Apple	In the set approximation of each a monitor state, which as for the				
	New case					
	ACCESSION COMP. But COMP.	APRIL OF TAXABLE PARTY	Content			
	60 million Recommende	1.1				
	ACURAL PORCESS	Manufacture FC	block			
	erace	THP STORAS				
	SubsTage Tale	Trachy Inc. Ballyne				
	Aftmals pipes to NE provide.	Paniers doos, hatra autorez.				
	- Advances in a new reard	- Data process many local local				
	- Jeliz is inter Brates area is forth Dia	- Date Mill Machine Millers and Miller				
	- and the second second	-				
straductidite wave OR.2 OR.50 OR.50 Sectors Sector Sector Sector						
SEARCH & man () more () more ()						
Execution of the second s						
NUMBER OF TAXABLE	and other paint in later	CARLINS, DUT DELTA DAMES				



#### Why there are so many duplicates?

C

- Same content, different URLs, often with different session IDs. User 1 Server User 2
- Crawling time difference



## Tradeoff of online vs. offline removal

	Online-dominating approach	Offline-dominating approach
Impact to offline	High precision Low recall Remove fewer duplicates	High precision High recall Remove most of duplicates Higher offline burden
Impact to online	More burden to online deduplication	Less burden to online deduplication
Impact to overall cost	Higher serving cost	Lower serving cost

## Software Infrastructure Support at Ask

- Programming support (multi-threading/exception • Handling, Hadoop MapReduce)
- Data stores for managing billions of objects Distributed hash tables, queues etc
- Hadoop Communication and data exchange among machines/services
- Execution environment
- Controllable (stop, pause, restart).
- Service registration and invoca
- service monitoring
- Logging and test framework.

## **Requirements for Data Repository** Support in Offline Systems

- Update
  - handling large volumes of modified documents
  - adding new content
- Random access
- request the content of a document based on its URL
- **Compression and large files**
- reducing storage requirements and efficient access
- Scan
  - Scan documents for text mining.

## **Options for Data Stores**

- Bigtable at Google
- Dynamo at Amazon
- Open source software

Apache Cassandra	Bigtable Dynamo	Java/Hadoop	Apache
Hypertable	Bigtable	C++/Hadoop	Baidu
Hbase	Bigtable	Java/Hadoop	Apache
LevelDB	Bigtable	C++	Google
MongoDB		C++	