الجـامـعـة الأهـلـيـة
AHLIA UNIVERSITY
BAHRAIN

An efficient approach in Near Duplicate Document Detection using Shingling-MD5

Your
global
future
begins
here

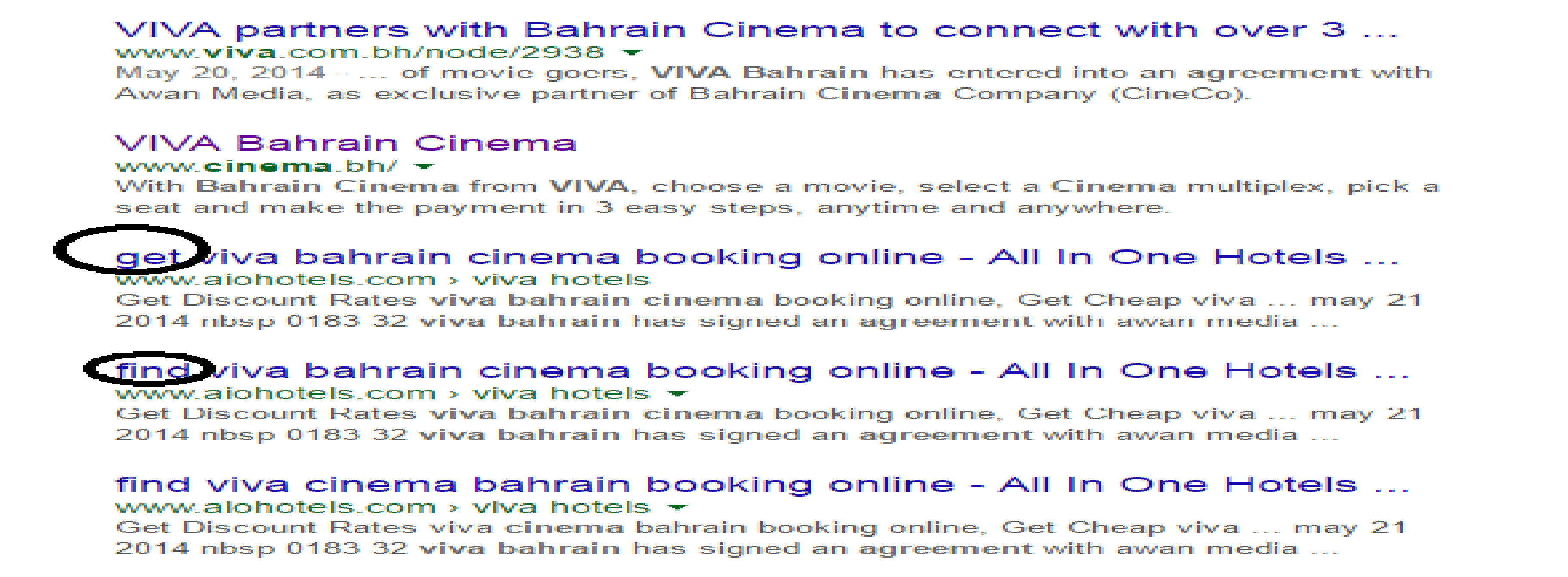
College of Information Technology

Dr. Suresh Subramanian

Problems in the web search

- Web is huge, diverse, and dynamic
- We are currently drowning in information and facing information overload.

Example : Viva Bahrain Cinema Agreement



VIVA partners with Bahrain Cinema to connect with over 3 ...
www.viva.com.bh/node/2938
May 20, 2014 - ... of movie-goers, VIVA Bahrain has entered into an agreement with Awan Media, as exclusive partner of Bahrain Cinema Company (CineCo).

VIVA Bahrain Cinema
www.cinema.bh/
With Bahrain Cinema from VIVA, choose a movie, select a Cinema multiplex, pick a seat and make the payment in 3 easy steps, anytime and anywhere.

get viva bahrain cinema booking online - All In One Hotels ...
www.aiohotels.com › viva hotels
Get Discount Rates viva bahrain cinema booking online, Get Cheap viva ... may 21 2014 nbsp 0183 32 viva bahrain has signed an agreement with awan media ...

find viva bahrain cinema booking online - All In One Hotels ...
www.aiohotels.com › viva hotels
Get Discount Rates viva bahrain cinema booking online, Get Cheap viva ... may 21 2014 nbsp 0183 32 viva bahrain has signed an agreement with awan media ...

find viva cinema bahrain booking online - All In One Hotels ...
www.aiohotels.com › viva hotels
Get Discount Rates viva cinema bahrain booking online, Get Cheap viva ... may 21 2014 nbsp 0183 32 viva bahrain has signed an agreement with awan media ...

Question Arises ?

- Users are forced to spend their effort, money and time unnecessarily during documents search on
  - Documents pertaining to user related query are retrieved partially
  - Requested documents retrieved are only partially relevant
  - Retrieved documents are not in order
  - **Duplication in retrieved documents**

(Haveliwala et al., 2002; LEE, 2007; Pohl et al., 2010, Dallal et al. 2012, Subramanian et al. 2014)

Problem Justification

Available methods are insufficient to reduce the duplication and retrieve most relevant documents according to user query (Wang et al., 2015; Aldallal et al., 2012; Dong, 2008; Picarougne et al., 2002)

* Quickly and efficiently determine which documents in a large set are similar to each other

* Identify Near duplicate documents which would improve the performance of a search engine to retrieve the documents without duplication

Related Work

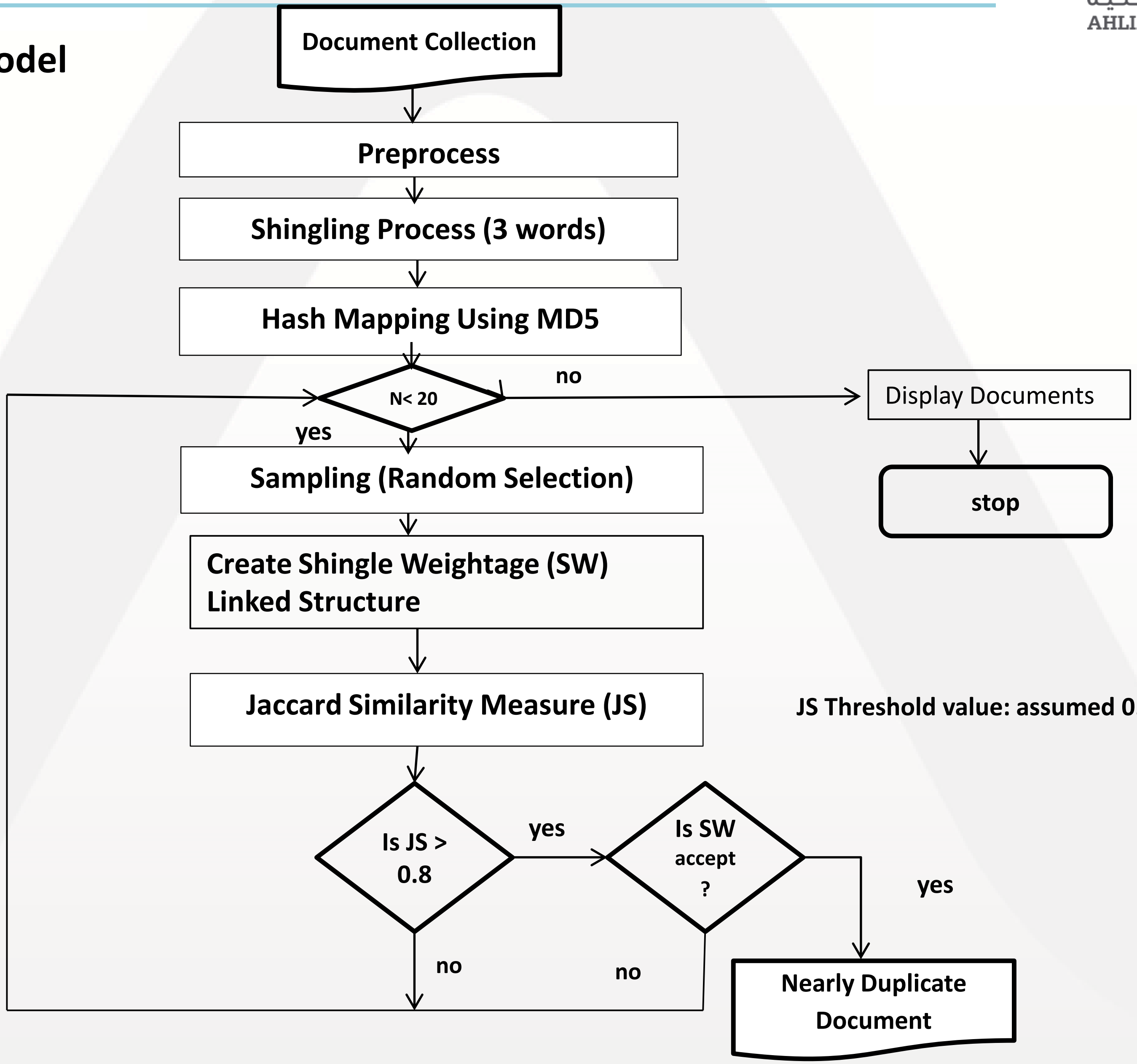
Duplicate data detection techniques

- divide the files into a number of parts
- compare corresponding parts between files via hash techniques
- matching the tags such as paragraph, anchor, heading tags etc..
- Identifying and Filtering Near-Duplicate Documents (Broader et al., 2000)
- Extract important phrase, or multi Word segments (Cooper et al., 2002)
- Near-duplicate Document Detection System with SIMD Technologies (Yuan et al., 2011)
- Adaptive near duplicate detection using similarity learning (Hajishirzi et al ., 2010)
- Word Weightage Based Approach for Document Detection for duplicate documents (Subrmanian et al., 2014)
- A fingerprint of Paragraph Generation Approach for Detecting Similar Document (Wang et al., 2014)

An efficient approach in Near Duplicate Document Detection using Shingling-MD5

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN

Proposed Model


Document Collection
Preprocess
Shingling Process (3 words)
Hash Mapping Using MD5
N< 20
no
yes
Display Documents
stop
Sampling (Random Selection)
Create Shingle Weightage (SW) Linked Structure
Jaccard Similarity Measure (JS)
Is JS > 0.8
yes
Is SW accept ?
yes
no
no
Nearly Duplicate Document

JS Threshold value: assumed 0.8

Similarity Measure

Step 1:

Preprocess

- Document Collection
- Stemming words
- Document collection with words (Linked Structure)

Step 2: Shingling Process

Example

D1 : I am Sam.

D2 : Sam I am.

D3 : I do not like green eggs and ham.

D4 : I do not like them, Sam I am.

If (k = 1) shingle of

D1 U D2 U D3 U D4 :

{[I], [am], [Sam], [do], [not], [like],[green], [eggs], [and], [ham], [them]}.

An efficient approach in Near Duplicate Document Detection using Shingling-MD5

<u>If (k=2) – shingles of</u>

{ [I am], [am Sam], [Sam I],
[am I], [I do], [do not], [not like], [like green], [green eggs], [eggs and], [and ham], [like them], [them Sam]}.

<u>If (k=3) – shingles of</u>

{ [I am Sam], [am Sam do], [Sam do not], [Sam not like], [am I like], …………

<u>4 – shingles or 5 shingles ….. ???</u>

Researchers proposed to have 3 shingles

An efficient approach in Near Duplicate Document Detection using Shingling-MD5

الجامعة الأهلية
AHLIA UNIVERSITY
BAHRAIN

Step 4: Find the hashing value for each shingle using MD5

MD5 – Is the popular hashing algorithm to create hashing value for the provided shingles

Doc_Id	Shingle Id	Shingles	Hash values
D1	S1	I am	4a4c38338
D1	S2	am Sam	d737avc93c
D2	S1	like green	34ue25rt93

Step 5: Continue until the number of specified number of times

Create sampling using random process

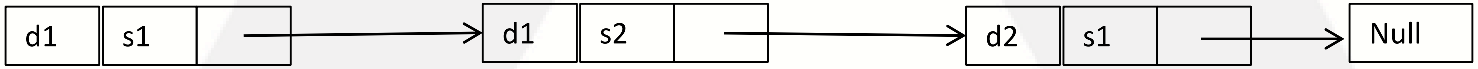
Randomly generate the number k; Where k between 1 and total number of shingles

An efficient approach in Near Duplicate Document Detection using Shingling-MD5

Step 6: Create linked data structure for Shingles document Id with weightage

Weightage calculated based on

* HTML tags such as <Title>, <Heading> and <A>



d1 s1 → d1 s2 → d2 s1 → Null

Step 7: Pick the Shingle, S_k from the list

While number of Document Shingles in the Sampling list :

Do Identify duplicates by comparing the Jaccard Similarity Index Created

How to calculate the similarity index using Jaccard Index

D1 : [I am], [am Sam]
D2 : [Sam I], [I am]
D3 : [I do], [do not], [not like], [like green], [green eggs], [eggs and], [and ham]
D4 : [I do], [do not], [not like], [like them], [them Sam], [Sam I], [I am]

Jaccard similarity (Sample)

JS(D1, D2) = 1/3 ≈ 0.333
JS(D1, D3) = 0 = 0.0
JS(D1, D4) = 1/8 = 0.125
JS(D2, D3) = 0 = 0.0
JS(D3, D4) = 2/7 ≈ 0.286
JS(D3, D4) = 3/11 ≈ 0.273

Jaccard Similarity:

$$JS(D1,D2) = \left|\frac{D1 \cap D2}{D1 \cup D2}\right|$$

An efficient approach in Near Duplicate Document Detection using Shingling-MD5

Step 8:

```
Check with threshold value (JS)
 if JS > 0.8 then  (0.8 is assumed)
          goto Step 9
 else
          goto Step 5
End
```

Step 9:

```
Check the documents Shingle weightage Summation (SW)
          if SW are equal or  acceptable (find the difference)
                    Store into Nearly Document Pool
                    goto Step 5
```

Step 10: Display nearly duplicated documents

An efficient approach in Near Duplicate Document Detection using Shingling-MD5

Implementation

- Implementation Using Java
- Document Set which can be collection of minimum of 100 documents
- Random duplication of documents

Results

-- Results shows that nearly duplicated documents have been identified

Future Work:

- Comparison has to be done with other standard models like Cosine, MinHash and SimHash model

- Any Questions ???
  - Further queries – please send mail to
    - ssubramanian@ahlia.edu.bh

Thank you