SURVEY ON FIXING THRESHOLD AND IMPLEMENTING EFFECTIVE DETECTION OF NEAR DUPLICATE WEB DOCUMENTS IN WEB CRAWLING

CPSC 8985 Fall 2015 Team – P1

By:
Manoj Reddy Bhimireddy,
Krishna Pavan Gandi,
Bhargav Roy Veeramachaneni,
Reuven Hicks,

Under the guidance of:
1/8/2016 Dr. Soon-Ok Park
AIM:

To fix the **Threshold** in the detection of duplicate and near-duplicate web documents in web crawling.
Introduction

- Web Mining is the branch of data mining that deals with the analysis of World Wide Web.
- The search engine technology has led to the development of World Wide.
- The search engines are the chief gateways for access of information in the web.
- The ability to locate contents of particular interest amidst a huge heap has turned businesses beneficial and productive.
Overall simplified Process

1. Web Index
2. HTML Document
3. Traverse links
4. Near-duplicate?
   - Yes: newly-crawled document(s)
   - No: insert
5. Entire Index
6. Trash

Date: 1/8/2016
Problem Definition

- Fixing the threshold to detect the web document is duplicate or not.
TEAM WORK

- Web Document Parsing
- Stop Words Removal
- Stemming Algorithm
- Keyword Representation
- SSM Value Calculation.
- Fixing threshold in Effective detection of Near Duplicate Web Document in web Crawling.
Solution Phase

- The documents are considered as near duplicates if its similarity scores are lesser than a threshold value.

- Memory for repositories has been reduced and the search engine quality has been improved owing to the detection.
DFD context level 0

Web DOC2

Process

Web DOC1

Ssm value to
detect whether
doc is duplicate
or not
DFD context level 1

Web Doc1

Web Doc Parsing ➔ Stop Word Removal ➔ Stemming algorithm

Web Doc2

Key Word representation ➔ Similarity calculation ➔ Ssm value

Duplicate or Not?
H/w & S/w Requirements:

- 80 Gb HDD
- 1 Gb RAM
- Pentium4 processor (or) Later
- WindowXP (OS)
- Java
- Action script 3.0
- MS Access

1/8/2016
ALGORITHM

Input(doc1, doc2)
Loop1(doc1)
{
    step1: web document parsing
    step2: stopword removal
    step3: stemming algorithm
    step4: keyword’s table formation
}

1/8/2016
Goto loop1(doc2)
step5: similarity score
step6: calculating SSM value.
step7: fixing threshold.
step8: comparing SSM with threshold.
step9: end
output(whether duplicate or not?)
MODULES

• Web document parsing
• Stop word removal
• Stemming algorithm
• Keyword’s table
• Similarity check
• Calculating SSM
Keywords Representation

<table>
<thead>
<tr>
<th></th>
<th>$K_1$</th>
<th>$K_2$</th>
<th>$K_4$</th>
<th>$K_5$</th>
<th>.....</th>
<th>$K_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>$C_1$</td>
<td>$C_2$</td>
<td>$C_4$</td>
<td>$C_5$</td>
<td>.....</td>
<td>$C_n$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$K_1$</th>
<th>$K_3$</th>
<th>$K_2$</th>
<th>$K_4$</th>
<th>.....</th>
<th>$K_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_2$</td>
<td>$C_1$</td>
<td>$C_3$</td>
<td>$C_2$</td>
<td>$C_4$</td>
<td>.....</td>
<td>$C_n$</td>
</tr>
</tbody>
</table>

$$a = \text{Index}(\left[ K_i \right]_{T_1})$$

$$b = \text{Index}(\left[ K_i \right]_{T_2})$$
Similarity score calculation

keywords present in both the tables

\[ S_{Dc} = \log(\text{count}(a) / \text{count}(b)) \times \text{abs}(1 + (a - b)) \]

\[ S_{Dc} = \text{Similarity Detection for Comparison table} \]

keywords present in T1 but not in T2 is taken as

\[ S_{DT1} = \log(\text{count}(a)) \times (1 + |T_2|) \]

\[ S_{DT1} = \text{Similarity Detection for T1 but not T2} \]

keywords present in T2 but not in T1 is taken as

\[ S_{DT2} = \log(\text{count}(b)) \times (1 + |T_1|) \]

\[ S_{DT2} = \text{Similarity Detection for T2 but not T1} \]
Similarity Score Measure

- The similarity score (SSM) of a page against another page is calculated by using the following equation.

\[ SS_M = \frac{\sum_{i=1}^{N_c} S_{Dc} + \sum_{i=1}^{N_{T1}} S_{DT1} + \sum_{i=1}^{N_{T2}} S_{DT2}}{N} \]

Where, \( N = \frac{|T1| + |T2|}{2} \)

- \( N_c \) = No of keywords in comparison table
- \( N_{T1} \) = No of keywords in \( S_{DT1} \) table
- \( N_{T2} \) = No of keywords in \( S_{DT2} \) table
Experimental values

Number of Documents

ssmvalue

1/8/2016
Actual Work

Things to be considered to fix the threshold are

- Timestamps
- Advertisement
- Counters
- SSM value

1/8/2016
Process

• Considering two identical web documents with same content but adding advertisements and timestamps.
• Calculating SSM value.
<table>
<thead>
<tr>
<th>Filename1</th>
<th>Filename2</th>
<th>ssmvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>14907</td>
<td>14907 - Copy</td>
<td>0</td>
</tr>
<tr>
<td>14907</td>
<td>14907 - Copy</td>
<td>16.1510079210364</td>
</tr>
<tr>
<td>14909</td>
<td>14909 - Copy</td>
<td>1.56362700836395</td>
</tr>
<tr>
<td>14909</td>
<td>14909 - Copy</td>
<td>1.56362700836395</td>
</tr>
<tr>
<td>14921</td>
<td>14921 - Copy</td>
<td>0.598452370046102</td>
</tr>
<tr>
<td>14921</td>
<td>14921 - Copy</td>
<td>0.598452370046102</td>
</tr>
<tr>
<td>14921</td>
<td>14921 - Copy</td>
<td>0.598452370046102</td>
</tr>
<tr>
<td>14921</td>
<td>14921 - Copy</td>
<td>0.598452370046102</td>
</tr>
<tr>
<td>14921</td>
<td>14921 - Copy</td>
<td>0.598452370046102</td>
</tr>
<tr>
<td>14907</td>
<td>14907</td>
<td>0</td>
</tr>
<tr>
<td>14909</td>
<td>14909 - Copy</td>
<td>16.4559285353099</td>
</tr>
<tr>
<td>14932</td>
<td>14932 - Copy</td>
<td>10.8803107225469</td>
</tr>
<tr>
<td>14941</td>
<td>14941 - Copy</td>
<td>8.6571278131262</td>
</tr>
<tr>
<td>14943</td>
<td>14943 - Copy</td>
<td>0</td>
</tr>
<tr>
<td>14949</td>
<td>14949 - Copy</td>
<td>4.52665933967439</td>
</tr>
<tr>
<td>14978</td>
<td>14978 - Copy</td>
<td>7.03328806521919</td>
</tr>
<tr>
<td>14982</td>
<td>14982 - Copy</td>
<td>1.99615466201221</td>
</tr>
<tr>
<td>15001</td>
<td>15001 - Copy</td>
<td>8.30652674718105</td>
</tr>
<tr>
<td>15004</td>
<td>15004 - Copy</td>
<td>3.5168563395619</td>
</tr>
<tr>
<td>14984</td>
<td>15484</td>
<td>51.7017793262597</td>
</tr>
<tr>
<td>15484</td>
<td>15531</td>
<td>0</td>
</tr>
<tr>
<td>14984</td>
<td>15484</td>
<td>0</td>
</tr>
<tr>
<td>14982</td>
<td>14982 - Copy</td>
<td>22.9784699038245</td>
</tr>
</tbody>
</table>
**GRAPH**

Number of Documents

ssmvalue

1/8/2016
Approximate Threshold

19.5043
Project Demo
MAIN FORM

Select First File
- Unknown Format
- Text Format
- Document Format

File 1

Select Second File
- Unknown Format
- Text Format
- Document Format

File 2

SSM Value
SRI LANKA GETS USDA APPROVAL FOR WHEAT PRO
Department of Agriculture approved the Continental C
Drs a tonne C and F from Pacific Northwest to Coloni
They said the shipment was for April 8 to 20 delive
REUTERS
SRI LANKA GETS USDA APPROVAL FOR WHEAT PRICING

Department of Agriculture approved the Continental Grain Co. offer of 6.5 dlrs a tonne C and F from Pacific Northwest to Colombo.

They said the shipment was for April 8 to 20 delivered.

REUTERS

CCC ACCEPTS EXPORT BID FOR WHEAT FLOUR T

Agriculture Department said

The bonus awarded was 113.0 dlrs per tonne and

The wheat flour is for delivery May 15-June 15, 1981.

An additional 162,500 tonnes of wheat flour are still

Reuter
SRI LANKA GETS USDA APPROVAL FOR WHEAT PRODUCTION

The Sri Lanka Department of Agriculture approved the Continental Grain Co's proposal to sell 300,000 tonnes of wheat to the US.

They said the shipment was for April 8 to 20 delivery.

REUTERS

CCC ACCEPTS EXPORT BID FOR WHEAT FLOUR T

The Canadian Wheat Board (CWB) said it has accepted a bid from CCC to export 162,500 tonnes of wheat flour.

An additional 162,500 tonnes of wheat flour are still available for export.

中的文

Sir: This button is Start Working or ALT+z Alt-Z

SSM Value 72.3017458674453


• www.eurojournals.com/ejsr_32_4_08.pdf
• ilpubs.stanford.edu:8090/831/1/2008-14.pdf
Questions or Doubts
Thank You