Reliability Verification of Search Engines’ Hit Counts
- How to select more reliable hit count within a SE -

Takuya FUNAHASHI and Hayato YAMANA
(yamana@acm.org)
Computer Science and Engineering Div.,
WASEDA University JAPAN
Have you ever experienced?

8,870 to 2,070?
Where have 6,800 web pages gone?

Put a Query and
Click “Search” Button

How can we have reliable hit counts?

Click “Next” Button

We want to make clear it.
Agenda:

**PART1:**
Importance of Search Engines’ hit counts

**PART2:**
Related Work

**PART3:**
Experiment and Trustworthiness of hit counts

**PART4:**
Conclusions
Introduction

1. IMPORTANCE OF SEARCH ENGINES’ HIT COUNTS
Background – How important hit-count is?

Many researches based on Hit Count exist …

### Researches using Hit Count

**Translation Support**


**Calculate Similarity between Words or Sentences**


**Evaluation**


NLP researches using web data are mostly based on SE’s hit counts

SE’s hit counts now become one of indispensable web resources
Example

by Y.Yamamoto et al.
(Fact Search System)

Original Proposition
The dinosaurs became extinct 65 million years ago

User selects verification target

The dinosaurs became extinct 65 million years ago

System splits sentence

Query
“The dinosaurs became extinct” & “years ago”
P1

Search Engine

Snippets
Returns 1000 results

… The dinosaurs became extinct 65 million years ago …
… The dinosaurs became extinct 55 million years ago …
… The dinosaurs became extinct 10,000 years ago …
… The dinosaurs became extinct years ago …

Alternative Terms
Extracts alternative terms

65 million 55 million 10,000

Queries
dinosaurs & became & extinct & “65 million years” & ago
dinosaurs & became & extinct & “55 million years” & ago
dinosaurs & became & extinct & “10,000 years” & ago

Search Engine

Collects web pages

Frequencies of Propositions
The dinosaurs became extinct 65 million years ago: 10000 pages
The dinosaurs became extinct 55 million years ago: 2000 pages
The dinosaurs became extinct 10,000 years ago: 10 pages

It utilizes the difference among hit counts.
Hit count v.s. the number of returned web documents

- Difference between the Hit Count on the 1\textsuperscript{st} SERP and the number of returned web documents

Experiment by using 1,000 queries on Nov. 2008
How often does the relationship turn over?

- Using 1,000 queries to compare the relationship, i.e. which has large hit count, for every pair of 1,000 queries

CASE1: comparing the hit counts on the 1st SERP.
CASE2: comparing the numbers of returned web documents

We should make clear the trustworthiness of search engines’ hit counts

turn over ratio of the number of results of every pair
2. RELATED WORK
Related work

• **Googleology is Bad Science**
  
  Queries repeated the following day gave hit counts varied by more than 10%.

• **Quantitative comparisons of search engine results**
  
  He compared three search engines, i.e. difference among the three.

• **Investigation of the accuracy of search engine hit counts**
  
  He compared three search engines to find out the search engines returning the most accurate hit count.
  He assumed the number of returned web documents is correct.
None of previous researches provide what the reliable hit count in a search engine is.
3. EXPERIMENT AND TRUSTWORTHINESS OF HIT COUNT
We cannot trust search engines’ hit count

Where have the 6,800 pages gone? Sometimes, hit count decreases to 1/10 or increases to 10 times.

Click “Next” Button
Purpose of Our Research

To provide “reliable hit count in a search engine” for researches using it

Cues of hit count ‘dance’ (change):

CASE1: Clicking the “Search” button many times

CASE2: Clicking the “Next” button step by step to reach the last page of the search results

CASE3: Searching with the same query on different days
Experimental Setting

Target:

- Google
- Yahoo!
- bing

(APIs)

API Settings:
- non-phrase search
- normal-safe filter
- Japanese language

Queries:
- 10,000 queries, provided by Yahoo! JAPAN as the top 10,000 frequent queries in December 2007.

Experimental Period:
- From October 2009 to December 2009.
Case 1: Clicking the “Search” button many times

1. Submit the same query to each SE 100 times in 5 min.
2. Calculate the coefficient of variation (CV) from 100 Hit Counts.

\[
CV = \frac{\text{standard deviation}}{\text{average}} = \sqrt{\frac{\text{variance}}{\text{average}}}
\]

3. Execute 1 and 2 for all queries, then produce its histogram.
Histogram of the coefficient of variation

<table>
<thead>
<tr>
<th>range</th>
<th>Google</th>
<th>Bing</th>
<th>Yahoo!</th>
</tr>
</thead>
<tbody>
<tr>
<td>cv = 0.0%</td>
<td>9,977</td>
<td>699</td>
<td>9,096</td>
</tr>
<tr>
<td>0.0% &lt; cv &lt;= 0.1%</td>
<td>9</td>
<td>2,555</td>
<td>730</td>
</tr>
<tr>
<td>0.1% &lt; cv &lt;= 0.5%</td>
<td>0</td>
<td>6,191</td>
<td>46</td>
</tr>
<tr>
<td>0.5% &lt; cv &lt;= 1%</td>
<td>0</td>
<td>171</td>
<td>4</td>
</tr>
<tr>
<td>1% &lt; cv &lt;= 5%</td>
<td>0</td>
<td>56</td>
<td>1</td>
</tr>
<tr>
<td>5% &lt; cv &lt;= 10%</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>10% &lt; cv &lt;= 20%</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>20% &lt; cv &lt;= 100%</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>100% &lt; cv</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sum</td>
<td>9,986</td>
<td>9,689</td>
<td>9,877</td>
</tr>
</tbody>
</table>

All Search Engines’ CV are almost less than 5.0%

We may ignore the dance in case 1.
Case 2: Clicking the “Next” button step by step

1. Submit one query to Search Engine to get HitCount(1,10), HitCount(11,20), ..., HitCount(991,1000)
   
   1st SERP         2nd SERP         100th SERP

2. Calculate Deep hit count Vector(DV)

   \[
   DV = \begin{bmatrix}
   \frac{\text{HitCount}(1,10)}{\text{HitCount}(1,10)} & \frac{\text{HitCount}(11,20)}{\text{HitCount}(1,10)} & \cdots & \frac{\text{HitCount}(991,1000)}{\text{HitCount}(1,10)}
   \end{bmatrix}
   \]

3. Execute 1 and 2 for all queries, then apply k-means clustering to a set of DVs

   note: Case 2 is applied to only Bing and Yahoo! because Google API does not return 1000 results.
Clustering (by using k-means)

• Objective
  – extract transition patterns of the hit counts

• Method
  – vary its clustering size $k$ from 1 to 6, then, select the best size by manual.
  – choose the best size based on the following points;
    1) **start offsets** when hit count dances begin are **clearly different** among clusters.
    2) **curves of change ratio** are **clearly different** among clusters.
Transition of Hit Counts - Case 2

Adjusting the final hit count to the number actually returned web documents.

Every 100 results, Yahoo! re-calculates the hit count.
At first, hit count should be calculated by using top-k results. When requesting more result, search engine will use another top-k results to calculate hit count.

Most reliable hit count will be the hit count when “k” is the largest number, but not adjusted to such as Bing.
Search Engines’ Pruned Index Architecture

At first, hit count is calculated by using top-k results when requesting more result. Reliable Hit Count in Case 2 is

\[ \text{HitCount}(k, k+9) \]

where \( k \) is defined as the largest number, usually 991 for a search having more than 1,000 matched pages. If a search engine adjusts the last hit count, we should use the hit count just before the adjusted count.
Case 3: Searching on different days

1. Submit queries to Search Engine every day during 2 months. 
   (From October 11, 2009, to December 12, 2009)

2. Calculate Vectors of Variational ratio (VV).

   \[ VV = \left\{ \frac{\text{HitCount}(10/11)}{\text{HitCount}(10/11)} , \frac{\text{HitCount}(10/12)}{\text{HitCount}(10/11)} , \ldots , \frac{\text{HitCount}(12/12)}{\text{HitCount}(10/11)} \right\} \]

3. Apply k-means clustering to VVs like Case 2.
Clustering Result - Case 3

(k=4) (k=5)

Ratios
Change Ratio
within a week
changes more than 30%

Search Date
change more than 30%
Clustering Result - Case 3

Change Ratio

Search Date

(k=3)

within a week
changes more than 30%
Conclusion of Case 3

Search Engines have two phases

“Dancing Hard Phase” : index update phase. During this phase, hit counts will change more than 30% within a week from our observation.

“Stable Phase” Hit counts do not change more than 30% within a week.

Reliable Hit Count in Case 3 is...

A hit count when a Search Engine is on stable phase.
4. CONCLUSIONS
Conclusion

- Hit count has become one of indispensable web resources
  - Reliable hit count can be the last hit count when a search engine is on stable phase where there exists small change less than 30% during one week.

- Our next challenge
  - Building some benchmark to correct/refine the hit count more reliable.
  - Providing some quality/reliable measure for the hit count.
THANK YOU FOR YOUR ATTENTION