Using Internet Search Engine Hits to Determine Truth Values

Shazril bin Azman[1] Azlan Iqbal[2] Azureen binti Azmi

[1, 2]College of Information Technology, Universiti Tenaga Nasional, Putrajaya Campus,
Jalan IKRAM-UNITEN, 43000 Kajang, Selangor
e-mail: shazril@uniten.edu.my, azlan@uniten.edu.my, azureenazmi@yahoo.com

***Abstract* - Search hits, or the number of hits returned by an internet search engine for a particular term refers to the estimated number of pages in the World Wide Web that contain the term. In this paper, we propose an extraction method to access the hit count and analyze the reliability of using it to determine the truth values of statements or terms to differentiate from erroneous ones. For example, this can be used to determine the correct spelling of places, to detect grammatical errors, and even for simple fact-checking. Our findings suggest a positive correlation between the number of hits and the 'correctness' of the search phrase.**

Keyword- search hits, hit count, external application, truth value

I. INTRODUCTION

The number of hits returned by a search engine for a search term or phrase is the estimated number of web pages found in the World Wide Web (WWW) that contain such a string of characters. A search engine accepts the search string from the user and displays a list of relevant websites along with the hit count inside the 'search engine result page' (SERP).

However, the search hit is not an exact result. It is just an estimation of how many relevant pages found by the search engine. You can google any keyword and keep clicking 'next' until you reach the end of the result page. There, you will notice the total numbers of pages are different, even lesser than the estimated hits. Nevertheless, the search hit is reliable in determining what we term as 'truth value', or the degree of factuality between true and false statements, which will be explained more in section III(C).

Currently there are more than one hundred search engines worldwide for different categories ranging from businesses, books, enterprises and games [1]. For example, *filecrop.com* provides searching for various files uploaded into popular file hosting services such as Rapidshare and Mediafire. The search engine in *jobstreet.com* provides information on job vacancies in Malaysia. A 'spider' or a 'crawler' is a program used by search engines to explore World Wide Web to retrieve hyperlinks to relevant web pages based on the keywords. The number of search hits is often displayed together with the search results.

This information can be used to determine the truth value of terms or statements. For example, we can determine the correct or *more* correct spelling of a word; "Amazon" or "Amahzon" because the former should have a higher number of search hits. We can also determine which one has correct grammar, for example, "it's all right" or "it's alright" and we can also determine whether a certain statement is factual or not, for example "earth has a round shape" or "earth has an oval shape", whereby the former should have higher number.

Section III explains how the search hit is extracted, along with the collected results presented in the subsection C with the line charts to support the truth value hypothesis.

II. LITERATURE REVIEW

Meng et al. in [2] proposed a technique to automatically extract the search hits for any search engine and any query, and they highlighted the importance of search hits in obtaining the document frequency of a term, estimating the size of the search engine and generating a search engine summary.

Fregtag in [3] pointed out that because WWW consists primarily of text, information extraction is central to any effort that would use the Web as a resource of knowledge discovery.The Honto Search system [4], prioritized the 'trustworthiness' aspect of information in determining whether a proposition or statement is true or false. Honto provides the user with popularity estimation of a phrase and its alternatives on the Web to ensure the trustworthiness of the information.

We used textual extraction approach to extract the search hits to determine truth values of statements or terms. We categorized 90 true and false statements,

keywords and terms, into three different comparative categories, which are general knowledge fact checking, spell checking and language and grammar checking. Higher scores of truth value imply more 'correctness'.

III. METHODOLOGY

Search engines typically use the HTML form to pass the query to the engine's server. The HTML form is used for providing inputs to the user. The inputs are the text fields, radio buttons, and checkbox. Usually, users have to key in the words or make selections in the Form. Any HTML design (form, font, background image, etc.) for any webpage is based on textual code. We can access this HTML code in any webpage in by right clicking and choosing 'view page source' in the web browser in any operating system. Usually the HTML code will start with '<html>' or '<!doctype html>' tag at the beginning of the code.

There are other languages used by web developers to access databases in the server, such as PHP (Hypertext Preprocessor) and ASP (Active Server Pages), but these languages will be converted into the HTML format when they are running. Therefore, in general, every webpage is in HTML format. In a sense, it is the 'machine code' of the Web.

When a user sends a query to the search engine, the list of hyperlinks for the relevant match of the query will be displayed in response. The resulting hit count can be obtained for external application through the extraction of the textual elements within that HTML source. In this research, the application to extract this value was developed using Microsoft Visual Basic 6.0 (VB6) and the Python programming languages. Visual Basic 6.0 provides the interface for the main controls such as clickable buttons, text boxes and labels. These objects will interact with the Python module to access the search engine and sends all the extracted information back to VB6 to be displayed. A single programming language would suffice but using two in this way served other purposes not directly related to this work.

Python provides a class and easy-to-use functions that are specifically dedicated to HTML extraction that fits the criteria needed in the research, but the Python application is an executable file and it is only console-based.

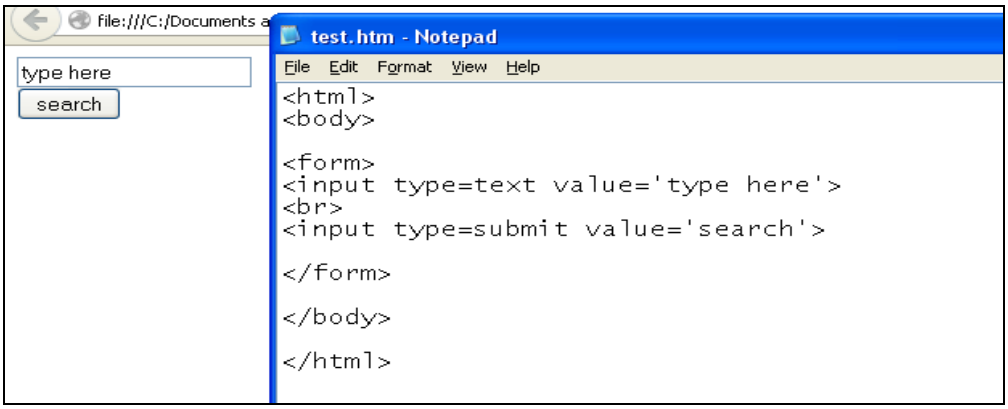


Figure 1. A simple HTML interface with its page source (right)

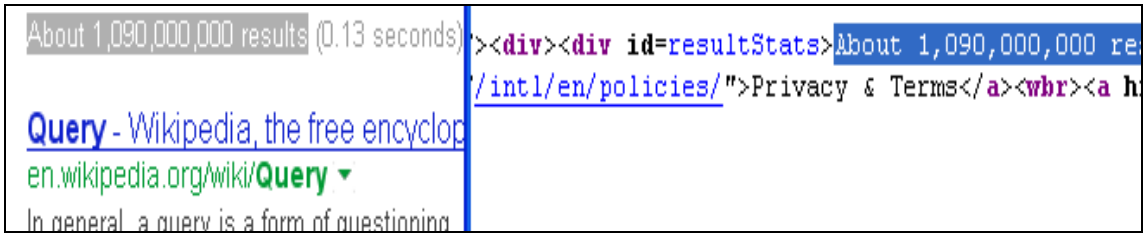


Figure 2. Google search result page (left) and its page source (right) with highlighted search hits

One Python downloadable class called *Beautiful Soup* enables Python to grab the HTML elements in the search engine result page (SERP). Beautiful Soup is a Python class that is especially made for reading through lengthy HTML code in Python and extracts certain elements in the code. This project uses this class to grab and parse the hit value to the VB6 application.

Python itself is an executable portable application. The whole Python project can be ported or backed up into another drive by copying and pasting the whole Python folder, but the total size is large, up to 100 MB. In order to reduce the size and not to use unnecessary libraries, the 'py2exe' [5] utility was used to make this Python program portable and it also provides only specific Python libraries for the application and thus, it makes the size smaller.

A. Hit Count Extraction

If a user types a keyword and it appears in the URL after the search button is clicked, this indicates the search engine is using the GET method. GET method is mostly used by common search engines. For example, if a user types 'Hello World', and clicks the search button, the same name will appear somewhere at the URL, indicating that the string is parsed through the GET method. The '&q= Hello+World&' is the string passed to the server when the user hits the search button. '&' is the variable's separator, 'q' is the name of the variable, and 'Hello+World' is the value/string of the variable from the textbox.

Figure 3. A URL's query string

This enables Python to send HTTP request via GET method to any desired search engine. Due to terms of service, certain search engines like Google does not allow any third party to access their page source via external application other than through the interface and the instructions that they provide [6]. Therefore, we decided to use Bing search engine because there is no such issue there presently.

Python's Beautiful Soup class enables the program to extract elements inside the HTML id tag and this makes the extraction efficient. Most web pages consist of 'element id', which gives a tag name to certain elements in the HTML. For example, in Google, the name for the element id of the search hits is 'resultStats', for Yahoo, it is 'resultCount' and for Bing it is 'count' as shown:

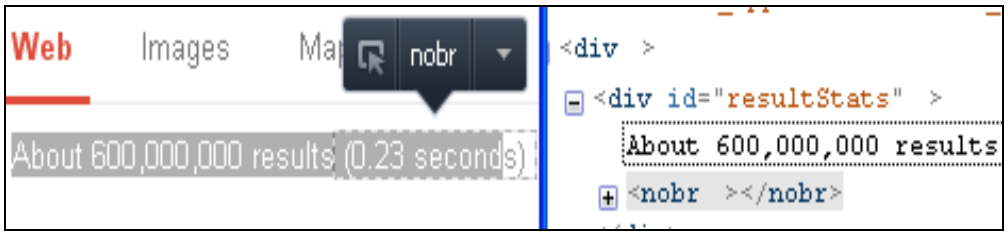


Figure 4. Google's hit result (left) and its source code (right)

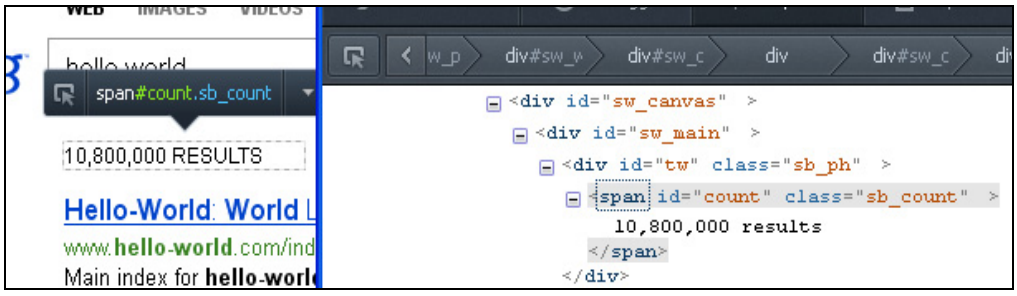


Figure 5. Bing's hit result (left) and its source code (right)

Figure 6. Yahoo's hit result webpage (top circle) and its source code (bottom circle).

B. Requesting HTTP for SERP

Python's 'urllib' library was used for sending HTTP request via GET method. Bing server will return SERP in response. The syntax goes:

>>urllib.urlopen("http://www.bing.com/search?q="+ whatUserType)

Then, all the SERP source code will be assigned to Python's Beautiful Soup variable as a string.

The hit value, which is contained inside 'Span' element, which is , will be extracted from that variable through the following syntax:

>>soup.find("span", id="count")

The resulting output from the syntax would be '10,800,000 results' as such. The comma delimiter and the word 'results' would be removed and the new values would be converted from string to integer, 10800000. The tested algorithm works as shown in the following flowchart:

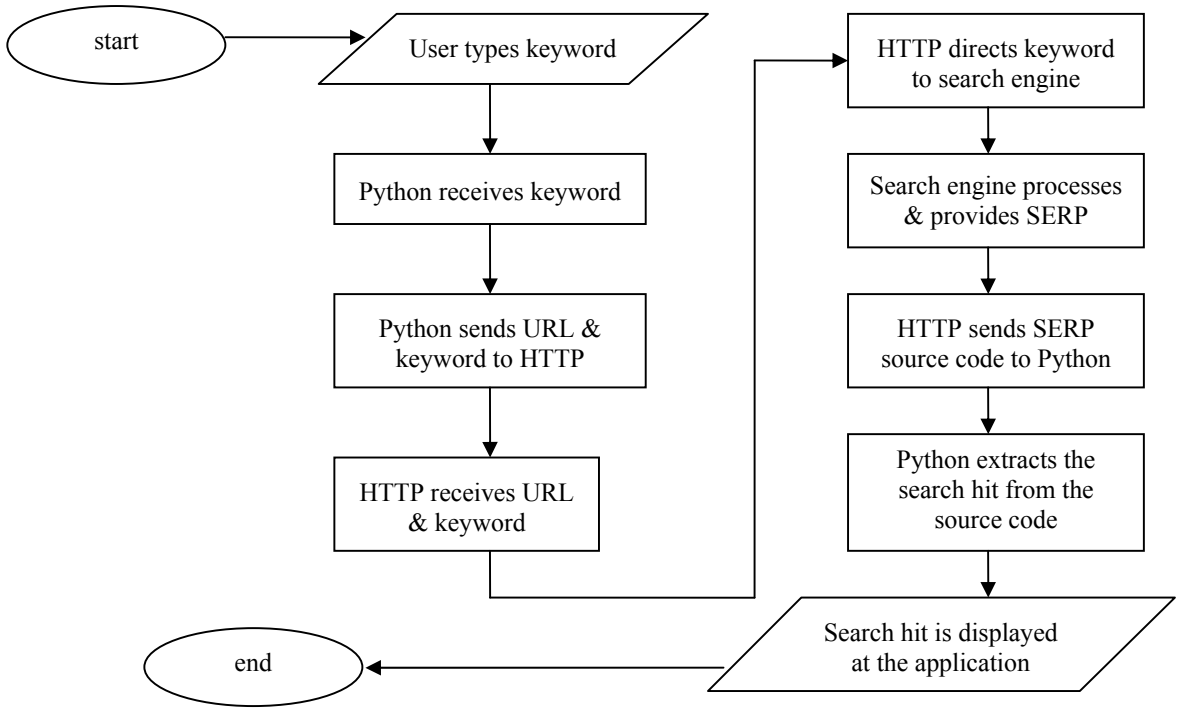


Figure 7. Flowchart of the external application at work

.

Figure 8. The Bing SERP and the Python-based external application at work

C. Information Accuracy

The search hits for correct keywords and erroneous keywords were compared. Three search engines (Google, Yahoo, Bing) were used for the comparison. The results show the correct keyword typically produced a higher number of hits. The following table shows the results for general knowledge fact checking (data collected in 2010):

Table 1. Results for general knowledge fact checking

Term 1	**"Barack Obama" "president of the United States"**	**"Barack Obama" "not president of the United States"**
Google	3 170 000 results	129 000 results
Yahoo	2 940 000 results	8070 results
Bing	6 480 000 results	8080 results
Average	4 196 666.67 results	48 383.33 results
Term 2	**"Venus" "is the hottest planet"**	**"Venus" "is not the hottest planet"**
Google	251 000 results	5410 results
Yahoo	19 700 results	335 results
Bing	17 500 results	230 results
Average	96066.67 results	1991.67 results
Term 3	**"Pacific" "is the largest ocean"**	**"Pacific" "is not the largest ocean"**
Google	935 000 results	2130 results
Yahoo	11 700 results	9 results
Bing	46 900 results	9 results
Average	331200 results	2148 results
Term 4	**"Mercury" " is the closest planet to the sun"**	**"Mercury" " is not the closest planet to the sun"**
Google	1 080 000 results	886 results
Yahoo	56 200 results	16 results
Bing	43 000 results	13 results
Average	393066.67 results	305 results
Term 5	**"Sahara " "is the largest desert"**	**"Sahara " "is not the largest desert"**
Google	518 000 results	3520 results
Yahoo	12 100 results	12 results
Bing	38 000 results	12 results
Average	189366.67 results	1181.33 results
Term 6	**"Nile" "is the longest river in the world"**	**"Nile" "is not the longest river in the world"**
Google	3 700 000 results	4980 results
Yahoo	38 000 results	16 results
Bing	29 600 results	54 results
Average	1255866.67 results	1683.33 results
Term 7	**"Bismarck" "is the founder of modern Germany"**	**"Bismark" "is not the founder of modern Germany"**
Google	14 000 results	483 results
Yahoo	16 results	8 results
Bing	16 results	8 results
Average	4677.33 results	166.33 results
Term 8	**"Tofu" "is made from soybeans"**	**"Tofu" "is not made from soybeans"**
Google	157 000 results	7 results
Yahoo	17 300 results	6 results
Bing	21 100 results	6 results
Average	65133.33 results	6.33 results
Term 9	**"Rafflesia" "is the biggest flower in the world"**	**"Rafflesia" "is the smallest flower in the world"**
Google	28700 results	1940 results
Yahoo	17600 results	115 results
Bing	369 results	88 results
Average	15556.33 results	714.33results
Term 10	**"Kent" "is known as the Garden of England"**	**"Kent" "is not known as the Garden of England"**
Google	167 000 results	13200 results
Yahoo	7170 results	12 results
Bing	7130 results	12 results
Average	60433 .33 results	4408 results

The following table shows result analysis from spell checking:

Table 2. Results for spell-checking

Term 1	**"Rendezvous"**	**"Rendezvoos"**
Google	41 200 000 results	9 180 results
Yahoo	39 900 000 results	128 results
Bing	36 700 000 results	142 results
Average	39 266 666.67 results	3 150 results
Term 2	**"Czechoslovakia"**	**"Checkoslovakia"**
Google	31 500 000 results	109 000 results
Yahoo	16 700 000 results	20 300 results
Bing	15 500 000 results	0 600 results
Average	21233333.33 results	56633.33 results
Term 3	**"Harvard University"**	**"Harverd niversity"**
Google	44 000 000 results	7 410 results
Yahoo	45 300 000 results	27 500 results
Bing	40 600 000 results	124 000 results
Average	43 300 000results	52 970 results
Term 4	**"Sacrilegious"**	**"Sacrelegious"**
Google	2 240 000 results	5740 results
Yahoo	4290 000 results	7200 results
Bing	4 060 000 results	483 results
Average	353 0 000 results	4474.33 results
Term 5	**"Mississippi"**	**"Mississipi"**
Google	403 000 000 results	4070 000 results
Yahoo	274 000 000 results	3560 000 results
Bing	208 000 000 results	5750 000 results
Average	295 000 000 results	4460000 results
Term 6	**"Fredericksburg"**	**"Fredericksburgh"**
Google	32 800 000 results	501 000 results
Yahoo	36 500 000 results	0 results
Bing	24 200 000 results	83 300 results
Average	31166666.67 results	194766.67 results
Term 7	**"Massachusetts"**	**"Massauchusetts"**
Google	647 000,000 results	13300 results
Yahoo	408 000 000 results	25000 results
Bing	38 700 000 results	59400 results
Average	364566666.67 results	72466.67 results
Term 8	**"Presbyterian"**	**"Presbaterian"**
Google	53 200 000 results	10 700 results
Yahoo	61 800 000 results	48 800 results
Bing	49 800 000 results	52 000 results
Average	54933333.33 results	37166.67 results
Term 9	**"Reykjavik"**	**"Reykavik"**
Google	30 200 results	270 000 results
Yahoo	7 340 000 results	21 800 results
Bing	15 500 000 results	48 000 results
Average	7623400 results	113266.67 results
Term 10	**"Wolfgang Amadeus Mozart"**	**"Wolfgang Amadius Mozart"**
Google	18 100 000 results	1230 results
Yahoo	5320 000 results	25 results
Bing	4900 000 results	17 results
Average	944 0000 results	424 results

The following table shows result analysis from language/grammar checking:

Table 3. Results for language/grammar checking

Term 1	**"a school of fish"**	**"a group of fish"**
Google	1 680 000 results	275 000 results
Yahoo	72 600 results	55 500 results
Bing	146 000 results	34 400 results
Average	1898600 results	121633.33 results
Term 2	**"a master's degree"**	**"a masters degree"**
Google	85 400 000 results	43 900 000 results
Yahoo	3990 000 results	6 710 000 results
Bing	9 880 000 results	5 860 000 results
Average	330 90000 results	150 523 333.33 results
Term 3	**"there are many boys"**	**"there are much boys"**
Google	2850 000 results	3260 results
Yahoo	11 000 results	4 results
Bing	10 900 results	4 results
Average	519233.33 results	1089.33 results
Term 4	**"an elephant"**	**"a elephant"**
Google	53 800 000 results	593 000 results
Yahoo	7 910 000 results	346 000 results
Bing	15 800 000 results	587 000 results
Average	7751 0000results	508666.67 results
Term 5	**"a little salt"**	**"a few salt"**
Google	11 800 000 results	331 000 results
Yahoo	410 000 results	8150 results
Bing	1110000 results	15 500 results
Average	4440000 results	118216.67 results
Term 6	**"went to the house yesterday"**	**"going to the house yesterday"**
Google	2 320 000 results	8 results
Yahoo	4220 results	9 results
Bing	4050 results	9 results
Average	776090 results	26 results
Term 7	**"I have a dream"**	**"I has a dream"**
Google	20 200 000 results	426 000 results
Yahoo	8710000 results	20500 results
Bing	7780000 results	24100 results
Average	1223 0 000 results	156866.67 results
Term 8	**"the door of the car"**	**"the car's door"**
Google	71300 000 results	521 000 results
Yahoo	49 100 results	8910 results

Bing	47400 results	25800 results
Average	23798833.33 results	185236.67 results
Term 9	**"an apple"**	**"a apple"**
Google	72 300 000 results	13 900 000 results
Yahoo	16 800 000 results	5420 000 results
Bing	37 100 000 results	8390 000 results
Average	42066666.67 results	9236666.67 results
Term 10	**"a cat"**	**"an cat"**
Google	101 000 000 results	79 300 results
Yahoo	30 600 000 results	285 000 results
Bing	56 200 000 results	436 000 results
Average	62 600 000 results	266766.67 results

The following line charts show the differences of search hits between true and false terms/statements in those three categories (general knowledge fact, spell check, and grammar). The *y* axis is the average hits from each search engine (Google, Yahoo, Bing) for every statement. The average values of the hits have been scaled down to logarithm 10 for better visualization at *y* axis. The *x* axis represents all the statements in the category accordingly to its number. For example, the one on "Venus" in Table 1 would be statement number 2.

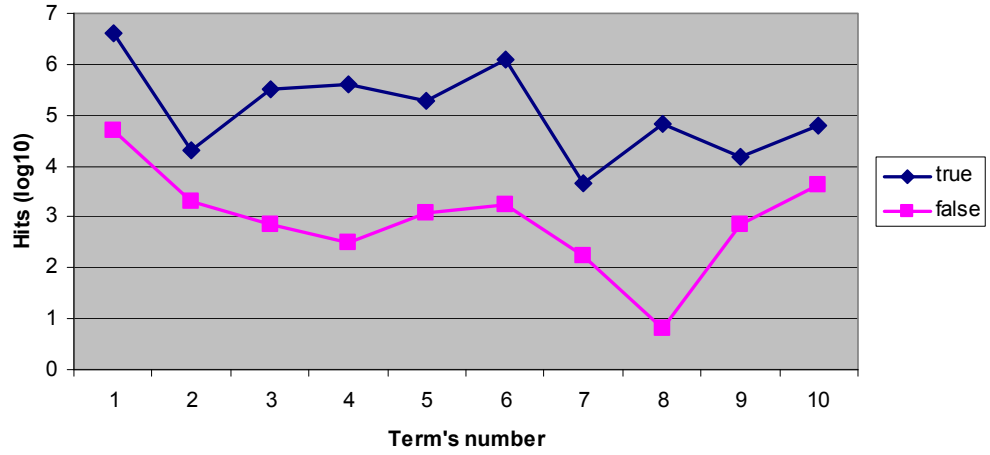


Figure 9. General Knowledge Fact Checking (average hit comparison between true and false terms/statements)

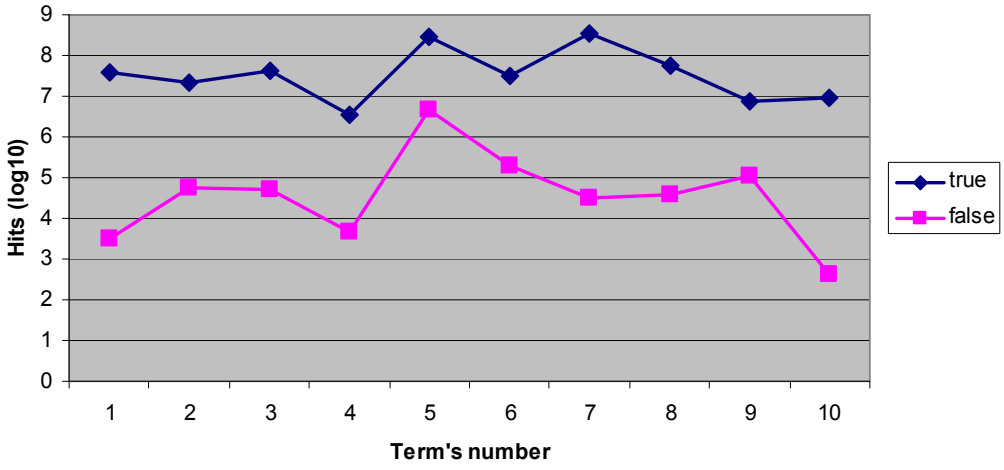


Figure 10. Spell Checking (average hit comparison between true and false terms/statements)

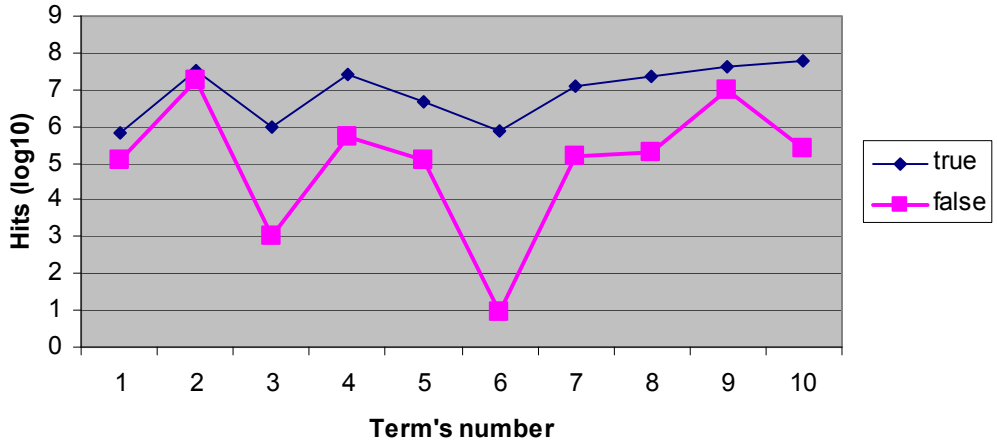


Figure 11. Spell Checking (average hit comparison between true and false terms/statements)

Occasionally, the number of the search hits differs from the previous reading. The search hit would 'dance' in some situations [7], for example if user searches for a term and he/she clicks the search button multiple times, or the same term is searched on a different day, the hit result would be different than the previous one.

The reason for this is unknown to us because all the technical detail behind the search engine is the company's intellectual property. However, to the best of our knowledge, this might be due to new data entry that contains the term that has been added or found by the engine's crawler. The followings are examples of the keywords and the different values of search hits obtained from Bing by multiple search clicks on June 5, 2013 and few screenshots of the application at work:

Table 4. The different results of same keywords obtained by multiple search clicks (Bing engine)

'Huntsman':	**'Altantuya':**	**'Snow White':**	**'Cloud Atlas':**
• 872,000 results	• 64,100 results	• 18,900,000 results	• 1,560,000 results
• 871,000 results	• 64,700 results	• 19,200,000 results	• 1,460,000 result
• 866,000 results	• 64,200 results	• 19,300,000 results	• 1,410,000 results

Figure 12. Comparison between "obama" and misspelled "obbama"

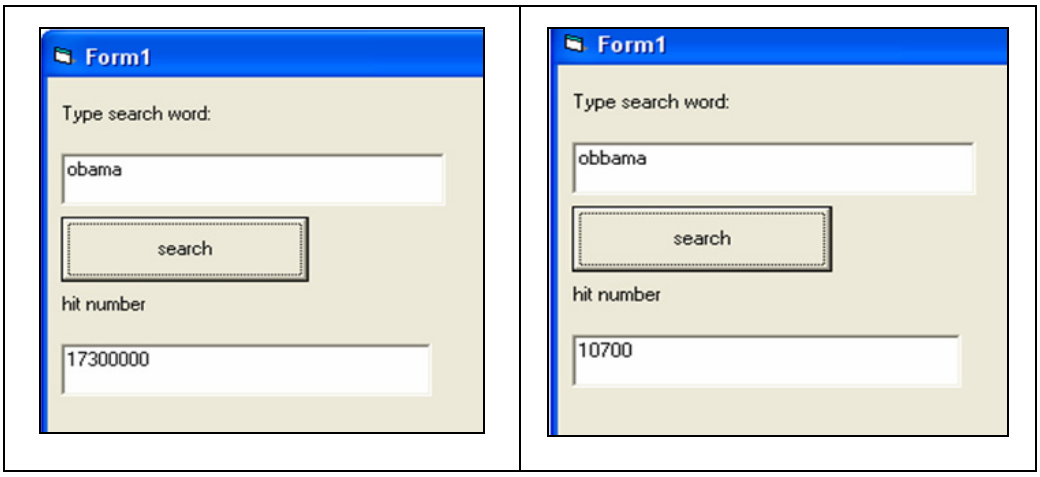
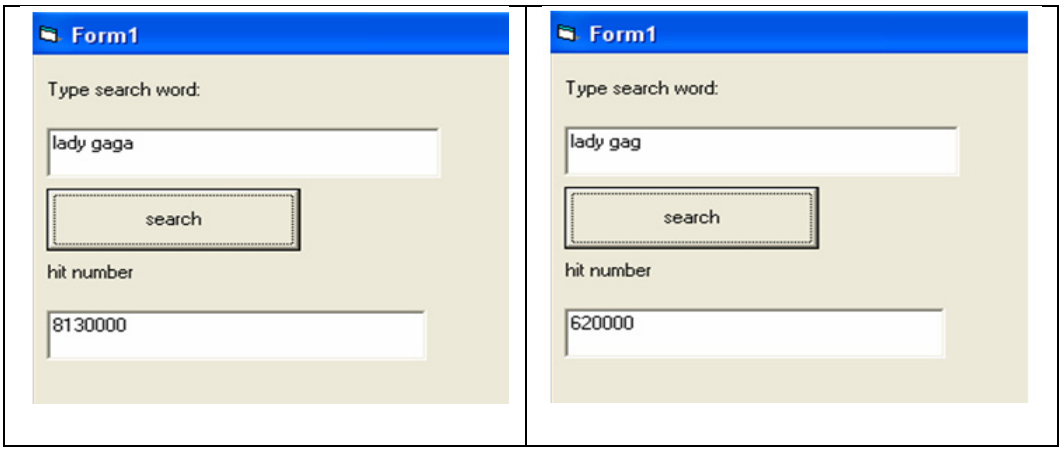


Figure 13. Comparison between “lady gaga” and misspelled “lady gag”



IV. CONCLUSION

In this article we have shown an extraction method of HTML hit results for the purpose of determining the truth value of search terms. This method can be useful to identify misspelled names of places, grammatical problems and also for simple fact-checking statements. While this may already have been known and implemented to some extent by popular search engines such as Yahoo!, Google and Bing, their precise techniques are typically industrial secrets. Users, however, may apply the information presented here to aid them in their daily tasks. A variety of third- party applications may also take advantage of search engines in this way.

V. ACKNOWLEDGEMENTS

We would like to thank the developers who provided us with open source tools that made this research possible. Special thanks to the Visual Basic and Python forum users who responded on various issues encountered. This research was sponsored in part by the Ministry of Science, Technology and Innovation (MOSTI) in Malaysia under their eScienceFund research grant (01-02-03-SF0240).

REFERENCES

[1] The search engine list. "The search engine list” Retrieved 16 May 2013, from http://www.thesearchenginelist.com/

[2] Y. Ling, X. Meng, W. Meng, “Automated Extraction of Hit Numbers from Search Result Pages” Retrieved 16 May 2013, from http://link.springer.com/chapter/10.1007/11775300_7#page-1

[3] F. Dayne.. “Information Extraction from HTML:Application of a General Machine Learning Approach”. 1998.

[4] Y. Yusuki, T. Taro, J. Adam, T. Katsumi. “Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis”. 2007

[5] Python py2exe. “py2exe” Retrieved 16 May 2013, from http://www.py2exe.org/

[6] Google policies and principles. “Google Terms of Service” Retrieved 16 May 2013, from http://www.google.com.my/intl/en/policies/terms/regional.html

[7] F. Takuya, Y. Hayato. “Reliability Verification of Search Engines’ Hit Counts”. 2010