Natural Language Web Application (NLWA) for Search Engine

Teh Phoey Lee and Tan Yee Teng,

UCSI University, Malaysia

Summary

The internet has become crowed as the website is growing increasingly which difficult for users to find the information they require without using the search engine. However, the current search engines are subject to the returned of irrelevant result due to the ambiguity problem. As the search engine are return the search result without consider the meaning of the word that user input. This paper proposed a prototype system named as Natural Language Web Application (NLWA) to response to the ambiguity problem that occurs in perform the search using the search engine. This system was implemented the fundamental concept of natural language processing (NLP) whereby is to differentiate the similar meaning (synonyms) or multiple meaning (polysemous) of the word if it has any. The NLWA is a middleware that link to Google search engine and able to generate the synonyms or differentiate the meanings of the word that input in NLWA and in turn return the result directly from Google search engine.

Key words:
Search Engine, ambiguity problem, NLWA.

1. Introduction

The size of the World Wide Web is growing increasingly which contained diversity of information like a library that full of resources for internet users to look for information. This phenomenon was provided both pros and con for the internet users. In terms of pros, internet users have more chances to access the variety sources of information that available from different websites over the World Wide Web. However, in terms of con, this increased amount of websites growth was lead to the difficulty for users to find the information they require. It is hardly to seek for information without using search engines unless having the specified URL for the particular web pages.

As a result, due to the World Wide Web phenomenon, web search engines were playing an important role in seeking information for the internet users. In this information explosion era, the search of information is highly relies on the search engines. Performing a search of Web to obtain desire information without using the search engines is a time consuming task. Therefore, search engines was became a common tool in looking for desire information from the range of web pages that growing progressively in the World Wide Web.

Today, internet users who access to the Web can see search engines everywhere regardless of it type or features. Based on Search Engine Watch.com, there is 319 million of searches performed everyday [1]. In spite of the common search engines browser such as Google, Yahoo!, Live Search, Ask.com and etc which provide a search for any information; there are numerous websites also contain a search engines which allow the search for contents in the particular own website. Certainly, most of the search engines have became standardized to provide a search for useful information. In this context, most of the search engines provide the search of web, images, video, news, blogs and even more.

However, the search engines are subject to the problem of generated unwanted or irrelevant information in the search result. The keyword(s) enter by user to perform a search may contain different meaning as represented different “sense” between verb and noun [6]. In other words, the keywords that input may have multiple meaning which can lead to ambiguity problem.

This ambiguity problem is mainly because of the search engines do not consider the exact meaning of the search query but only consider the keywords matching of the search query based on the indexes [6, 3]. In addition, the “keyword may not be able to convey complex search semantics a user wishes to express” [4] and thus return the irrelevant information that are not the user desires.

Due to the ambiguity problem, a prototype that applies the fundamental concept of natural language processing (NLP) was proposed. The idea of NLWA is to differentiate the meaning behind a word and to clustering the similar meaning of a word that input to the search engine. NLWA will function on top of Google search engines as a middleware to look into the meaning of the word that enters by user and in turn performs the search in Google search engine.

Manuscript received October 5, 2008
Manuscript revised October 20, 2008

2. How the Search Engine work

Search engines can differ dramatically in the way they find and index the documents on the Web, and the way they search the indexes from the user's query [5]. There are various types of search engines in the market and each type of it has their own working architecture. The two general type of search engines are discussed in this section. In this context, there are known as spider/crawler based and directory based [7].

There are three modules in the search engines architecture that work on spider or crawler based. First of all, in order for crawler based search engines to search the web and return information that looking by the user, search engines rely on a software robot which known as web spider or crawler [11, 12]. The web spider or crawler is a program that used to conduct a web crawling process. The web crawling process is to finds web pages or web documents from "site to site" [8] by following every links contained in each web page of a websites [10, 11, 12].

Subsequently, all the web pages that have been found or known as "crawled" by the web spider will be sent to indexing software in order to produce a list of indexes that based on the words in the content of each web page. The indexes that produced will then store in the search engine database. Some search engines may just crawl the title of the web pages whereas some may crawl every word in the web pages. Besides that, the indexes that produced may be vary depends on different search engines that using different of weighting [10, 8, 11].

Last but no least, when user performs a search, the search engines will match the word in user's search query with all the indexes that stored in its database [9]. Then, which in turn generate and return the search result from millions of the web pages according to the relevancy that determine by the search engine. Figure 1 illustrates the process of how the search engines work.

The major factors that determine the relevancy of search result is based on the algorithm to weigh the frequency of keywords appear and which part the keyword does appear in a web page as well as the popularity of web page [2, 13]. However, different search engines are using different algorithm to rank the search result. Also, the algorithm calculation of a particular search engine is a highly secret. Thus, this is also the reason of different search engines are provided different search result for the same search query [13].

Conversely, directory based search engines do not crawl the web pages but rely on human editors to work on the indexes and categorize the web pages into appropriate categories. In order for directory based search engine to return the information that search by user, the human editors match the search word with the description that submitted by the websites owner or the description review that written by the human editors themselves. In addition, the changes on the web pages would not be update automatically by the directory based search engines unless the description is update [15, 14].

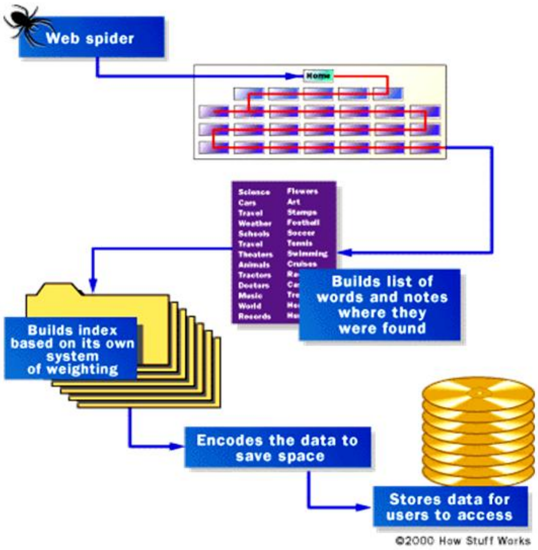


Fig 1. The process of how search engines work

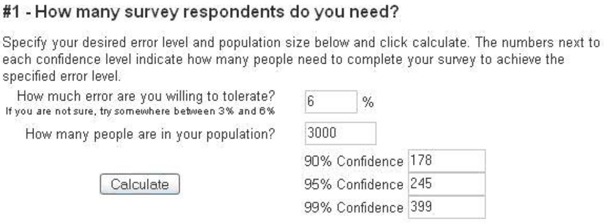
(Source: http://computer.howstuffworks.com/search-engine1.htm)

Yet accordingly, search engines are indexed large portion of web page which can lead to the problem of generated unwanted or irrelevant information. Dr. Bridge showed that "matching keywords from the user's query with words in the document may be prone to retrieving too many unwanted documents" [3]. It is clear that search engines only match with the words contained in user's query but do not consider the meaning of words as for a certain same word may have different multiple meaning, which is known as "polysemous" [6]. Therefore, this can also lead to the ambiguity problem whereby search engines just go to the keyword directly with web page that matches with the stored indexes [2, 11]. If the word that user search have multiple meaning, they may not interested with a search result that consist all meaning its have if they only interested in one of the meaning [11].

3. Questionnaire Distribution

Questionnaires are distributed based on the calculated sampling size on the population among UCSI students and the residents near the researcher's living area. The aim of this questionnaire distribution is to find out the top 10 used search engine among the users and their opinions on the returned result that gathered from the search engine.

The sampling size that is too large may lead to the waste of time and money, if is too small may lead to inaccurate result. Thus, the calculation of sampling size that used by researcher is based on the sampling size calculator as shown in the figure below;



#1 - How many survey respondents do you need?

Specify your desired error level and population size below and click calculate. The numbers next to each confidence level indicate how many people need to complete your survey to achieve the specified error level.

How much error are you willing to tolerate? 6 %
If you are not sure, try somewhere between 3% and 6%

How many people are in your population? 3000

90% Confidence 178
95% Confidence 245
99% Confidence 399

Figure 3.2 Sample Size Calculators by Custominsight.com [16]

Figure 2 indicate the sampling size is 245 respondents will be needed with the confidence level at 95%, confidence interval at 6% and population at 3000. The confidence level is the level of confidence of the sample answer that represent the true answer of the population. The confidence level is fall between 90%, 95% and 99%. The higher the confidence level, the higher the accuracy of the answer based on the population. Researchers select confidence level at 95%. Besides, confidence interval, also known as sampling error is the plus and minus figures in percentage that researcher willing to tolerate. In this case, researcher is using 6% as the maximum of error that willing to tolerate. Thus, with the confidence level at 95% and confidence interval at 6%, the sample is being asked between 89% (95% - 6%) and 101% (95% + 6%) of the total population. Lastly, the population is the total size of target population where the sample is selected based on 50% of UCSI students and 50% of public respondent from researcher living area which is about 3000.

The respondent's questionnaire is then collected for the process of analysis. The tool that use in the project to generate the statistic from the questionnaire result is Microsoft Office Excel 2003. Some of the results are presented on the next part of this paper.

4. Questionnaire Results Analysis

This section analyzes the result of the returned questionnaire of 300 copies that were equally distributed to 50% of UCSI students and 50% of public respondents.

4.1 Demographic

Table 1 illustrates the age group of target respondents in range of maximum, minimum, and means with 42%, 1%, and 31% respectively.

Table 1: Demographic

Age group	Max	Min	Means
	21-25	<15	16-20
Gender	Male	Female	
	49%	51%	
Categories	Student	Non-Exec	Exec
	50%	17.5%	32.5%
Online experiences	Max	Min	Means
	5 years	< 1 years	4 years

In terms of gender, it shows that the survey has conducted approximately equaled on male and female from both students and public respondents. The female students hold 27% whereas the public female hold 24%. On the other hand, the male students hold 23% whereas the public male holds 26%. In term of categories, there are both equally at 50% of the total of questionnaires that distributed whereby the public respondents have categories of 32.5% from executive level and 17.5% from non-executive level. Also, the online experiences of the target respondents are 40%, 11%, and 17% in the range of maximum, minimum, and means, respectively.

4.2 Top 10 Most Favorite Search Engine

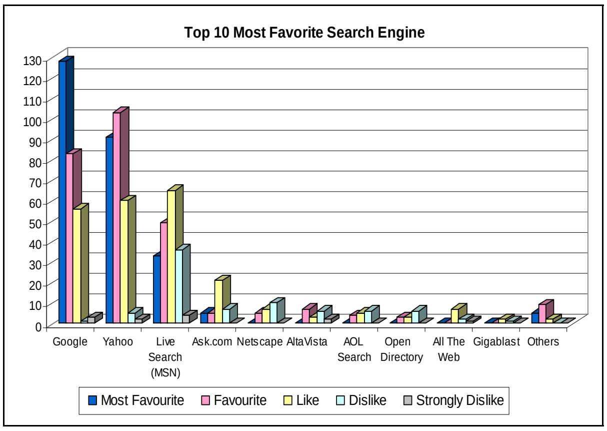


Fig. 2 Top 10 most favorite search engine used

Figure 2 shows the top 10 most favorite search engine used. Googles and Yahoo each hold the top on most favorite search engine chosen among the users while the gigablast are among the last two strongly dislike search engine being used. The choice of user's selection somehow reflected on the popularity of the search engine. However, researchers try to ask the respondent on the relevancy of the return of the search result and the analysis were done on Figure 3.

4.3 Top 10 Most Favorite Search Engine

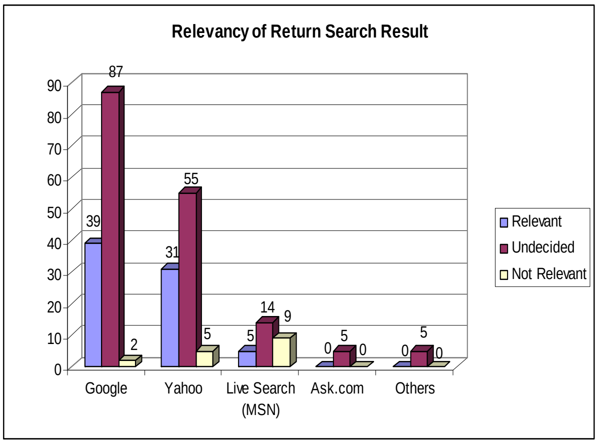


Fig. 3 Mostly favorite search engine with the relevancy of return search result.

As shown in Figure 3, Google is the search engine that has the highest relevant return of search results whereas Live Search is the highest irrelevant return of search results. However, the highest selection of the relevancy of return search result is “Undecided” regardless which of the most favorite search engine. Undecided refers as the relevancy is depends on what the respondents are search. Hence, based on the statistics of finding, the relevancy of the return of search results are strongly depends on the word that user search.

4.4 Top 10 Most Favorite Search Engine

Word that has multiple meaning may increase the chances of return irrelevant result. As according to Figure 4, there are 53% of respondents’ responded that the search for word that has multiple meaning was lead to the return of irrelevant result. It has proved the hypothesis of research whereby the ambiguity word may lead to the return of irrelevant result.

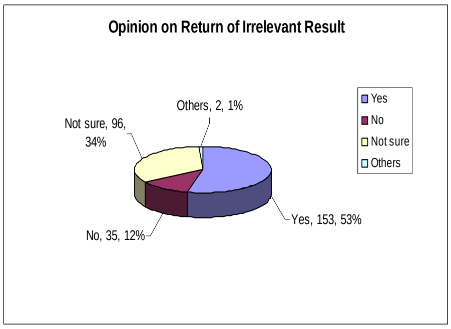


Fig. 4 Opinion on the reason of return of irrelevancy result.

5. Natural Language Web Application (NLWA)

NLWA work as the middle engine, try to cluster the similar meaning of word and in return on the search in search engines by differentiates between the types of word. The differentiation is done by looking into the type of word respective in verb, noun, and adjective to obtain the meaning behind the word.

The purpose of differentiate is to disambiguate the word that has multiple meaning. This is because certain word has multiple meaning in different type either is verb or noun which present different of meaning may lead to the retuned of irrelevant search results. As the meaning that present for the word that is polysemous are normally totally different. Additionally, the purpose of clustering is to group the similar meaning or synonymy of the word together in order to increase the chances of getting the information that user desire by looking into the meaning and the type of the particular word.

5.1 Framework

Figure 5 illustrates the framework of how the entire NLWA works in high level view whereby the system acts as a middle engine that function on the Google search engines to identify verb or noun or similar meaning of the word before perform the search. The identification is possible by clustering similar meaning of word in working with dictionary data based on thesaurus that store in the database.

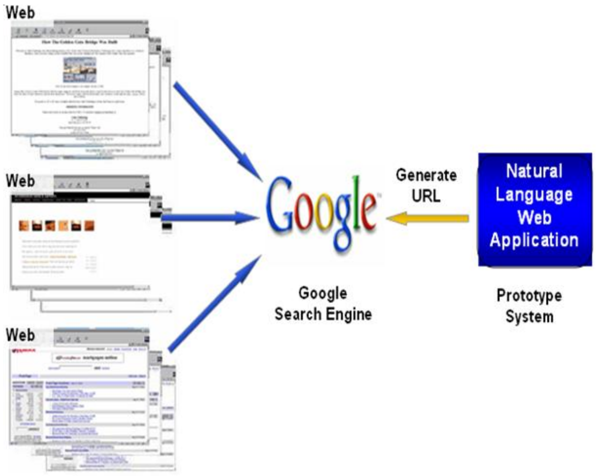
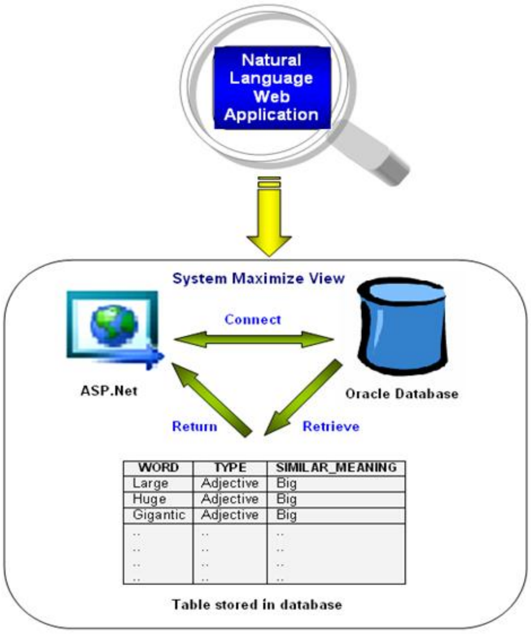


Fig. 5 Concept of NLWA

After the clustering of the similar meaning of word by looking into the type and the meaning behind of the particular word, NLWA will generate URL to connect to the Google search engine. NLWA is generating URL by passing parameter to inform the Google search engine what to search. The generation of URL will in turn link to Google search engines to perform the search. Subsequently, the Google search engines will return the search results according to the parameter in the URL it received.

5.2 Internal Process of NLWA

Figure 6 shows the specific view of the flow of internal process from the high level view system framework. The ASP.Net will connect to the database and retrieve the dictionary data from the table that stored in the database. After retrieved the data, it then return the value in order to generate the URL. Specifically, during the retrieving process, the ASP.Net will read row by row to look for the data that stored in the table "WORD" column. In this case, when its match the data that stored in the "WORD" column based on the word that user key in, it will return the value that stored in the "SIMILAR_MEANING" column to ASP.Net. The value refers to the similar meaning of word which is clustered in the database according to the meaning of the particular word. Lastly, when the value returns to ASP.Net, it will pass in each returned value as parameter to generate with the URL together in order to link to Google. With the parameters that pass it, Google is being informed what to search.



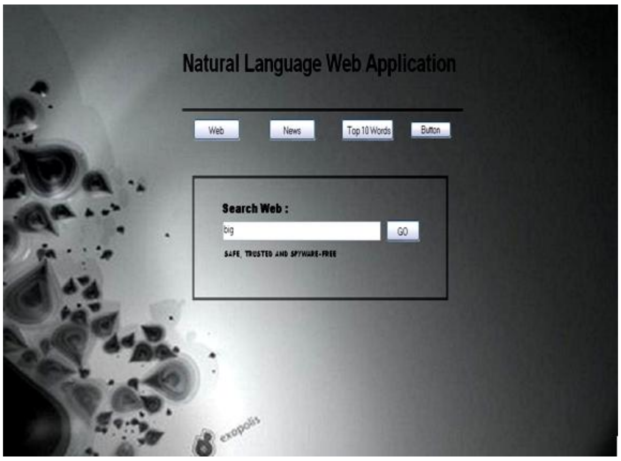
WORD	TYPE	SIMILAR_MEANING
Large	Adjective	Big
Huge	Adjective	Big
Gigantic	Adjective	Big
..	..	..
..	..	..
..	..	..
..	..	..

Table stored in database

Fig. 6 Framework of NLWA

6. Beta Testing of NLWA

In this section, a test for a group of 50 selected words is done in order to evaluate the meaning behind the word for the clustering process. For instances, Figure 7 shows the input of a word 'big'. User key in a word "big" as the input, NLWA will differentiate the types and meaning of "big". As a result, the input word has similar meaning and based on the clustering of its similar meaning, NLWA return "huge", "large", "gigantic" as the parameter.

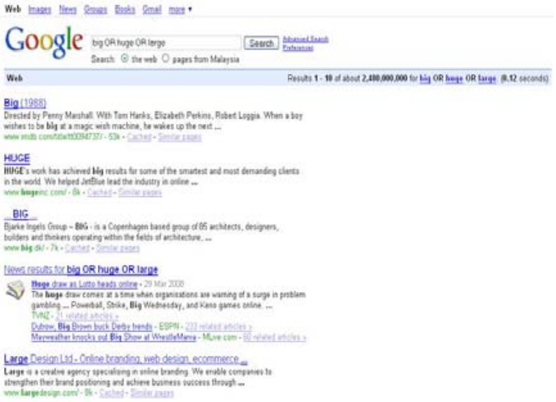

Natural Language Web Application

Search Web : big

SAFE, TRUSTED AND SPYWARE-FREE

Fig. 7 Testing page on input of keyword 'big' at NLWA

Subsequently, follow by the generation of the URL with the parameter that returned to trigger Google search engine to search and return the result for big, huge, large, and gigantic as shows in Figure 8.



Web Images News Groups Books Gmail more ▾
Google
Search
Advanced Search
Preferences
Search: ● the web ○ pages from Malaysia
Web

big OR huge OR large

Results 1 - 10 of about 2,480,000,000 for big OR huge OR large. (0.12 seconds)

Big (1988)
Directed by Penny Marshall. With Tom Hanks, Elizabeth Perkins, Robert Loggia. When a boy wishes to be big at a magic wish machine, he wakes up the next ...
www.imdb.com/title/tt0094737/ - 53k - Cached - Similar pages

HUGE
HUGE's work has achieved big results for some of the smartest and most demanding clients in the world. We helped JetBlue lead the industry in online ...
www.hugeinc.com/ - 8k - Cached - Similar pages

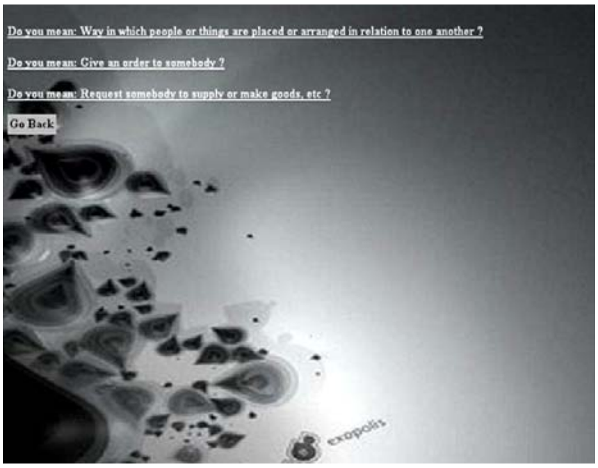
BIG
Bjarke Ingels Group – BIG - is a Copenhagen based group of 85 architects, designers, builders and thinkers operating within the fields of architecture, ...
www.big.dk/ - 7k - Cached - Similar pages

News results for big OR huge OR large
Huge draw as Lotto heads online - 29 Mar 2008
The huge draw comes at a time when organisations are warning of a surge in problem gambling ... Powerball, Strike, Big Wednesday, and Kino games online. ...
TVNZ - 21 related articles »
Dutrow, Big Brown buck Derby trends - ESPN - 233 related articles »
Mayweather knocks out Big Show at WrestleMania - MLive.com - 60 related articles »

Large Design Ltd - Online branding, web design, ecommerce ...
Large is a creative agency specialising in online branding. We enable companies to strengthen their brand positioning and achieve business success through ...
www.largedesign.com/ - 8k - Cached - Similar pages

Fig. 8 The return of result with the synonyms of big

Instead of clustering word that has similar meaning, NLWA differentiated word that has multiple meaning as well. Again, the input of a word "order" to the NLWA, the differentiation between the types and meaning of the word "order" is processed. As a result, the input word has multiple meaning in this case, NLWA then generates all the meanings of "order' as shows in Figure 9. User can select which meaning of "order" he or she is interested to perform a search. With the clustering of the similar meaning, Google search engine returned the search result according to the parameter on the meaning that user select. By showing all the meanings of the word that is polysemous, users can review which sense of meaning that the word present and select which they want to search.



Do you mean: Way in which people or things are placed or arranged in relation to one another ?

Do you mean: Give an order to somebody ?

Do you mean: Request somebody to supply or make goods, etc ?

Fig 9. Input of keyword 'order'

As a result, in comparing the results gathered from NLWA to the normal search on Googles, researchers find that the results returned applying the concept of NLWA has improved and also solved the ambiguity problem that occurs in perform the search using the search engine.

7. Conclusions

This paper has presented the how search engines work and explored the architecture of each. Secondly, based on the questionnaire analysis, the research hypothesis has been proved that the ambiguity word may lead to the return of irrelevant search result of the search engines. Moreover, based on the ambiguity problem of a word that is polysemous, the solution is designed to identify the meaning of word or synonyms for the word that input in the Natural Language Web Application (NLWA) are working. With the dictionary data which clustering for the similar meaning of word that stored in the database, the prototype is able to return either the meaning or synonyms of the word that input based on the type of the word in verb, noun, or adjective. Also, the test shows that the NLWA consists of the function that described and able to return the appropriate parameters.

8. Future Recommendations

As there are the limitations in the functionality of the prototype system, thus there are several considerations for the improvements of the system in the future as the list in following:

- Further study of searching for multiple words due to the scope of this prototype system is only to support in searching for single word.
- Enhance by further study of other approaches on natural languages processing in order to support the search in languages other than English language. For instance is Chinese, Malay and Japanese.
- Develop more attractive user interface and support for more features such as the link to not only Google search engine but other search engine as well. For instance are Yahoo, Live Search (MSN), and even others.

9. References

[1] University of California, Berkeley (2003), "How much information? 2003: Internet", http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm, 12 March 2008.

[2] Mary D. Taffet (2002), "Application of Natural Language Processing Techniques to Enhance Web-Based Retrieval of Genealogical Data", Syracuse University,

http://www.fht.byu.edu/prev_workshops/workshop02/fht2002paper27.pdf, 15 February 2008.

[3] Dr. Derek Bridge (2004), "Natural Language Processing (NLP)", University College Cork, Ireland, www.cs.ucc.ie/~dgb/courses/ai/notes/notes41.pdf, 18 February 2008.

[4] Qiang Yang, Hai-Feng Wang, Ji-Rong Wen, Gao Zhang, Ye Lu1, Kai-Fu Lee, Hong-Jiang Zhang, (2000) "Towards A Next-Generation Search Engine", pp. 1.

[5] TechWeb Network, "Search Engines", http://www.techweb.com/encyclopedia/defineterm.jhtml?term=Search+Engine , 24 February 2008.

[6] Jorge Gracia, Raquel Trillo, Mauricio Espinoza, Eduardo Mena, July 2006, 'Querying the web: a multiontology disambiguation method', ACM Press, University of Zaragoza, 22 February 2008, pp.241-242.

[7] Danny Sullivan (c) (2007), "How Search Engines Work", Search Engine Watch, http://searchenginewatch.com/showPage.html?page=2168031, 18 March 2008.

[8] Yelena Shapiro and Etelka Lehoczky (2003), "How do search engines work?", SearchEnignes.com, http://www.searchengines.com/search_engines_101.html, 25 February 2008.

[9] Elizabeth Liddy (2001), "How a Search Engine Works ", Director of the Center for Natural Language Processing Professor, School of Information Studies, Syracuse University, Vol. 9 No. 5, http://www.infotoday.com/searcher/may01/liddy.htm, 26 February 2008.

[10] UC Berkeley Library (Jan, 2008), "How do search engines work?", http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html, 25 February 2008.

[11] Curt Franklin, "How Internet Search Engines Work", http://www.howstuffworks.com/search-engine.htm, 26 February 2008.

[12] Wendy Boswell, "How Do Search Engines Work?", http://websearch.about.com/od/enginesanddirectories/a/searchengine.htm, 19 February 2008.

[13] Danny Sullivan (a) (2007), "How Search Engines Rank Web Pages", Search Engine Watch, http://searchenginewatch.com/showPage.html?page=2167961, 25 February 2008.

[14] Barbara Arena (2001), "The Complete Idiot's Guide to Making Money With Your Hobby", Alpha Books, ISBN 0028638255, p.80, http://books.google.com.my/books?id=q5aL3BGpN0YC&pg=PA80&lpg=PA80&dq=how+search+directory+work&source=web&ots=5V3t4FSOOy&sig=xHvI0YLdZ91XGObkM_lumr-2FSo&hl=en#PPA80,M1, 15 March 2008.

[15] Marckini (2000), "Types of Search Engines", http://www.inc.com/articles/2000/04/18599.html, 22 March 2008.

[16] Custominsight.com, "Sample Size Calculator", http://www.custominsight.com/articles/random-sample-calculator.asp 22 March 2008.

Teh Phoey Lee received the Bsc. in Remote Sensing from Universiti Technologi Malaysia in 2002 and M.S degrees in Information Technology from Universiti Sains Malaysia in year 2003. She started to pursue her Phd at University Putra Malaysia since 2007. She is now a lecturer at UCSI University. Her research areas include Information System and Software Engineering.

Tan Yee Teng is a student from UCSI University and will be receiving her B.Sc. (Hons.) Business Information Systems from UCSI in July 2009.