Platform for extraction, visualization and analysis of search trends

Abdul Wahid
Department of Software Engineering
Sofia University St Kl. Ohridski
Sofia, Bulgaria
awahid@gmail.com

Boyan Bontchev
Department of Software Engineering
Sofia University St Kl. Ohridski
Sofia, Bulgaria
bbontchev@fmi.uni-sofia,bg

ABSTRACT

Currently available web search engines and social media websites stick to their specific web traffic trends and do not provide generic overview of the whole web. We can better understand the current market and user interest about a specific product or topic or any term by analyzing all the trends from search engines and social media. Unfortunately, there is no such single platform which provides a market analysis tool which is based on combined trends of the entire web. In this paper, we have discussed our first prototype based on Google and Twitter search trends. In future, we intend to include more search engines and social media websites. Also semantics, opportunity recognition by implementing decision support system and intelligence features would be added to the system.

Keywords

Search Trends, web application, entrepreneurship, opportunity recognition

1. INTRODUCTION

The search engines and social media websites provide Search Trends (ST). Search Trends are the most frequently searched user queries [26] in the search engines and, in social media, they represent the most frequently talked topics. In other words it is information about "What people are searching for or talking about". For example, in the season of FIFA World Cup the one of the obvious search trend was "world cup" this is because most of the people were searching the information related to "world cup".

Google is a search engine [3] whereas Twitter is a social media site which is used for micro blogging [27]. Google logs the searches made by people and then over a specific period of time it calculates the most searched keywords. Google has a specialized web interface which provides the facilities for people to know what the people are searching on Google [32]. This can be used to analyze the user interest and know the current market.

In Twitter the most mentioned keyword in the status message of people's over a specific period of time becomes the trendy topic. Google and Twitter both provide the means to understand user interests on the web by analyzing their searches and status messages.

A market analysis tool ST Tool was proposed [23] which would collect the most popular queries and information from major search engines and social media websites to facilitate the process of idea generation through assisting the business executives and management in the visualization/analysis of the market and gap/opportunity identification. This tool can assist in market analysis and understanding the user behavior over the internet. Analysis based on this tool would be more reliable and accurate and will present a better picture of the web because it would extract the data from various sources. Apart from that it would be able to suggest topics for authors to write about by telling what people are mostly reading on the internet.

Our first prototype is based on one search engine "Google," and one social media site "Twitter". In future we are going to extend this prototype to build a complete commercialized product.

2. RELEATED WORKS

Searching over the internet is growing exponentially with the passage of time. There are many search engines [1] and on these search engines billions of search queries are executed each day. Search engines process millions of queries in seconds. Different search engine have different mechanism for processing user queries and displaying records. The page rank algorithm developed by Larry Page [2] is used by Google [3] to order pages after each user query. Yahoo, AOL, Ask and other on the other hand use different techniques.

Another research [4] about Alta Vista search engine states that Alta vista keeps a query log about queries which consists of a timestamp, cookie, query term, submission information and the submitter information. So we see that both Google and Alta Vista use different mechanisms for querying searches. There is another type of queries that can be referred as crawling technique and it is used to pull pages from millions of web servers [2].

FIT'10, December 21–23, 2010, Islamabad, Pakistan.

Many researchers are working towards extracting the trends in various search engines and social media websites to facilitate the people with search trends. Mika Kaki and Anne Aula discussed their experiences related to complexity in search engines [5]. Yair Shimshoni, Erin Efron and Yossi Matias introduced many interesting tools which explore what users are searching for; for example, Google Trends and Google Insight [6]. Ioannis Anagnostopoulos proved that when users search behavior is examined along with the ability of the Internet search services. It results in an effective meta-search [7]. It is observed that the social connections and mechanisms are the motives behind entrepreneurial activities [8]. A travel related searches study reveals that the social media appear in search engine results [9].

3. RESEARCH WORK

Integration of different data is always a cumbersome process especially when you try to integrate totally different systems. Google is a search engine whereas Twitter is a social media site which is based upon user short status messages. By examining the status massages of users on Twitter one can understand the interest of that user or community intention [27]. The most mentioned keyword in the status messages of the peoples over a specific period of time becomes a trendy topic. Google logs the searches made by people and then over a specific period of time it calculates the most searched keywords. Google [28] [29] [30] has a specialized web interface which provides the facilities for people to know what the people are searching on Google. This can be used to analyze the user interest and know the current market. However the Google does not cover the whole web thus any analysis made from Google result would not be reflecting the complete picture of the web.

If we want to analyze properly the user interest then we cannot ignore the fact that the web has different type of sources from where we can have more information about the user interests. There are two main types of sources that can be used to see the user interests on the web: first Search engines and, next, Social media. The main question that arises here involves the following issues:

- How to extract the data from search engines and social media websites?
- How to combine the data from search engine and social media websites?

In this study we constructed a prototype to extract, analyze, combine and display the trends from one famous search engine - "Google", and one famous social media portal - "Twitter".

3.1 Comparison of Search Trends

We used SWOT analysis techniques to compare search engine and social network search trends. SWOT analysis is the study of strengths, weaknesses, opportunities and threats about a product or tool. The SWOT analysis of Google [10] and Yahoo [11] Facebook [12] and Twitter [12] and the top trends of Facebook of 2009 [13] and the top trends of Twitter [14] were analyzed to understand the similarities and differences between them.

3.2 Comparison with ST Tools

To perceive a clear picture table 1 provides a comparison of the constructed ST Tool with other existing tools.

Table 1 Comparison of ST tool with existing tools

Tool Name	Google trends	Twitter trends	Trend graph	Choose keywords	Select dates	Trend comp-arison
Trendsbuzz	Yes	Yes	No	Yes	Yes	No
Trendsmap	No	Yes	Yes	No	No	No
Trendee	Yes	Yes	No	No	No	No
TwiGoogle	Yes	Yes	No	No	No	No
Twitter search	No	Yes	No	Yes	Yes	No
Trendistic	No	Yes	Yes	Yes	Yes	Yes
Twpopular	No	Yes	Yes	Yes	Yes	No
Whatthetrend	No	Yes	Yes	Yes	Yes	Yes
Twitscoope	No	Yes	Yes	Yes	Yes	No
Google trends	Yes	No	Yes	Yes	Yes	Yes
ST tool	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

The results represented in Table 1 show that our solution has a definite edge over all the existing tools because it offers everything at one place plus combining the results to have more better and precise analytical tool than currently available in the market.

4. SYSTEM ANALYSIS AND DESIGN

The software methodology that we have used to develop the ST Tool was iterative and incremental development of prototypes. However, this study is restricted only to the first prototype. The main types of requirements that relate to technical management [19] are:

4.1 Requirements for First Prototype

The first prototype which serves as a proof of concept, our requirements are very basic and simple enough to implement. They are listed as follows:

- Search trends should be extracted from Google and Twitter only
- User should have option to select the date and see the trends from previous dates;
- The search trends should be displayed in detail and combined form;
- User should be able to see the trends by giving the keywords.
- User should be able to compare two or more search trends.

4.2 Design and Architecture

There have been performed object-oriented analysis and design works relying on UML (Unified Modelling Language) as a mean

for describing the target system with the help of use cases, class diagrams, sequence diagrams and a components' diagram.

The use cases specify how the system is going to be used. The use case diagram clarifies the user requirements and helps in understanding these requirements.

Figure 1 specifies the use cases for our prototype.

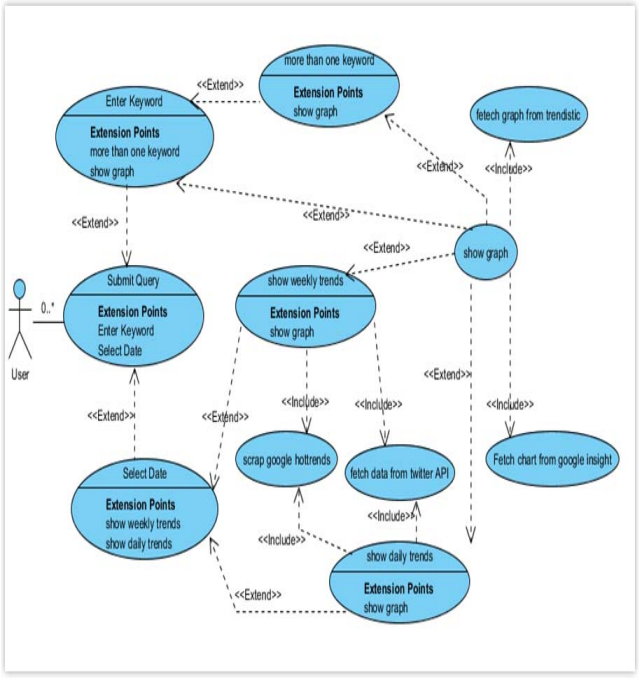

Figure 1 Use Case Diagram

Figure 2 shows the class diagram which specifies the classes, their attributes and relationships between the classes to describe the structure of a system.

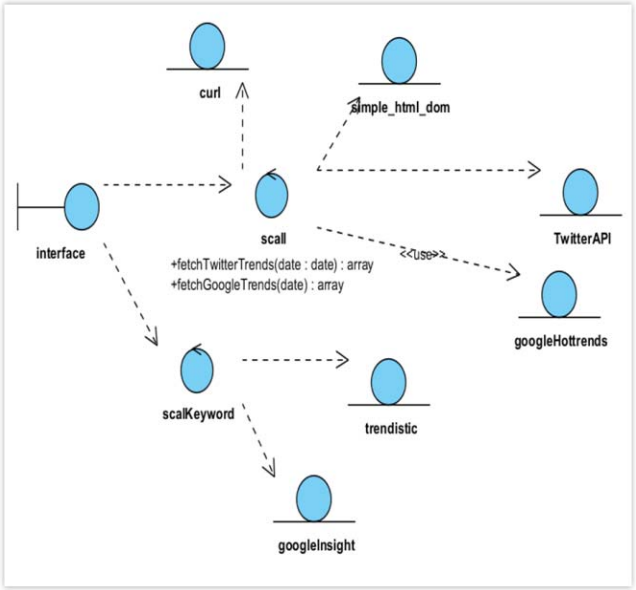

Figure 2 Class Diagram

In Figure 3 the "enter keyword" use case behaviour is shown.

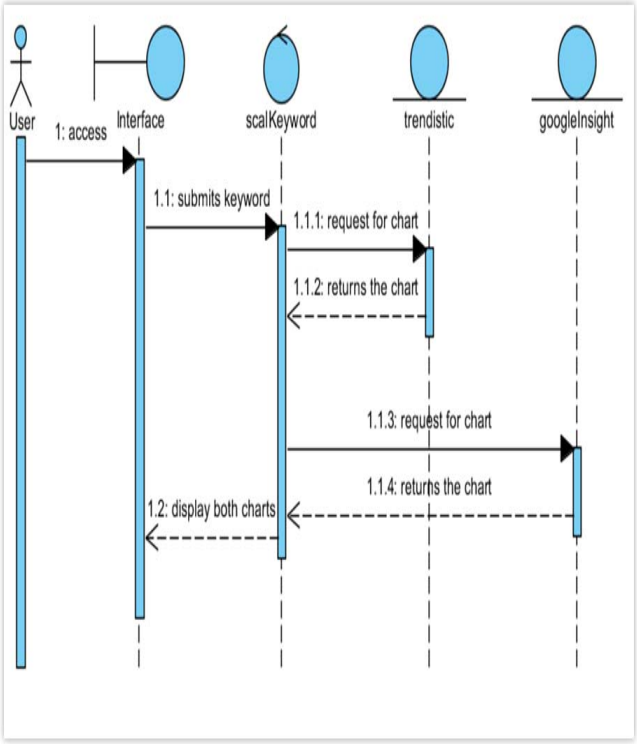

Figure 3 Sequence Diagram - Enter Keyword

Figure 4 shows the deployment diagram of the system.

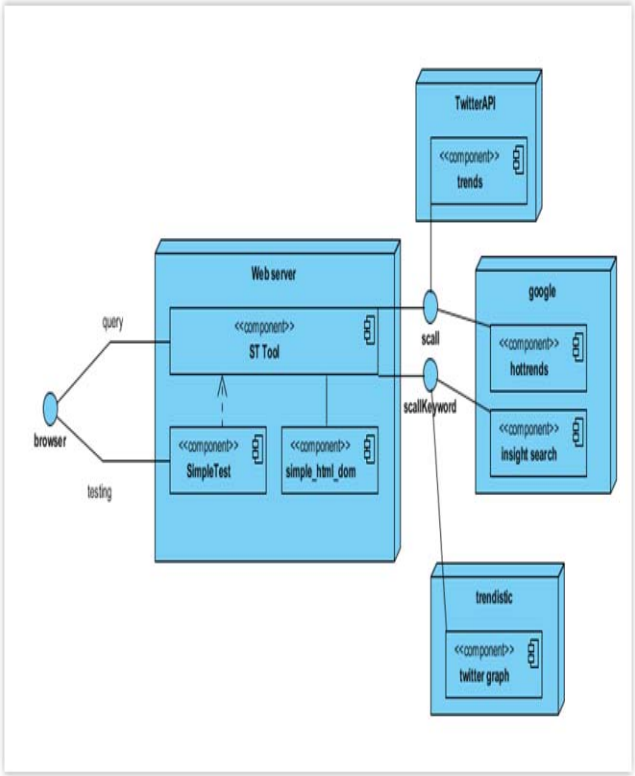

Figure 4 Deployment Diagram

4.3 User Interface

Figure 5 and 7 are the graphical representation of the search trends from Twitter and Google respectively. Figure 8 shows the detail search trends from Twitter and Google and figure 8 shows the combine search trends from both.

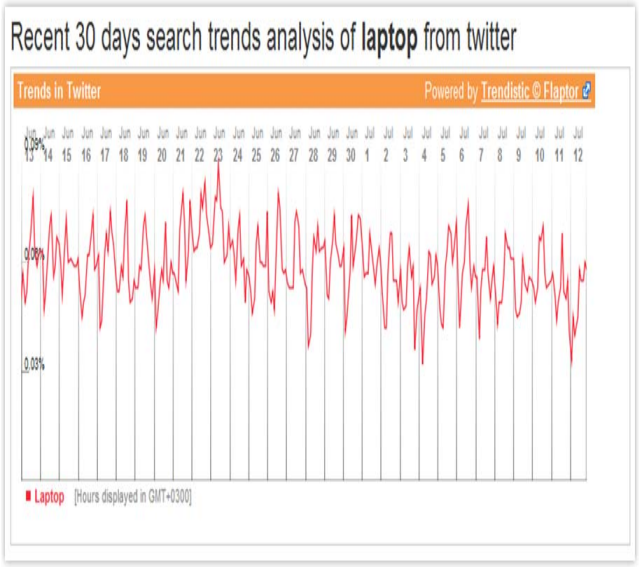

Figure 5 Trendistic Chart [18]

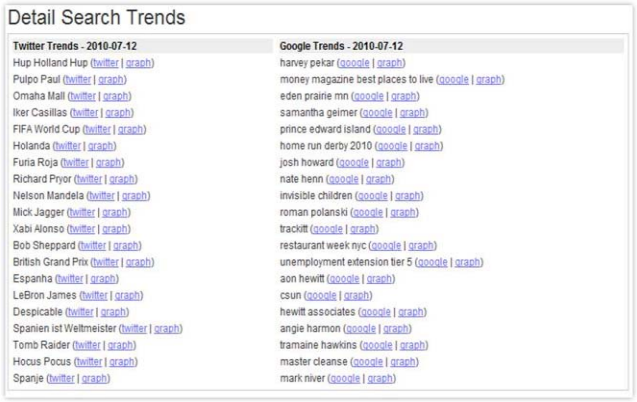

Figure 6 Detail Trends

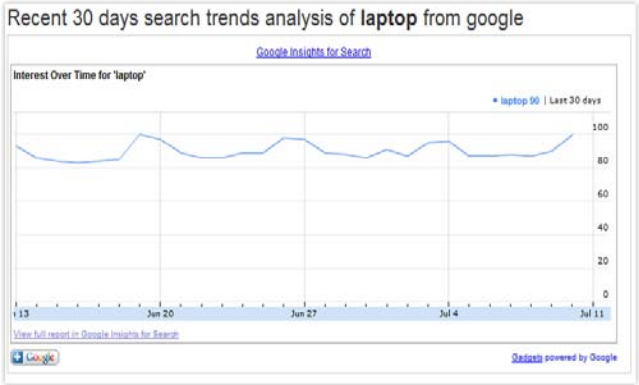

Figure 7 Google Insight Chart

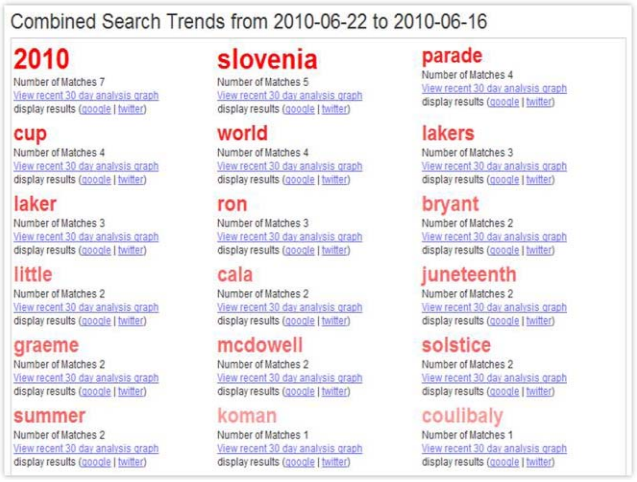

Figure 8 Combine Search Trends

In implementation of our first prototype, one of the top web servers – Apache [25], PHP as software application platform and Zend [15] for development were used.

5. Testing

Our prototype is developed in PHP and there are various automated testing software's available in the market to test the systems. Two well known tools for testing are PHPUnit and SimpleTest [20] which is able to check the page navigation, cookie testing and form submission [22]. PHPUnit [21] only works for PHP 4 and cannot be run from browser. We used SimpleTest for testing our prototype along with manual testing. We performed following testing on the system

- Functionality Testing
- Usability testing
- Interface testing
- Compatibility testing
- Performance testing
- Security testing

Functional testing was performed by SimpleTest and manually by checking the functionalities of the prototype from the browser. The usability testing was performed by conducting the survey which contained the five questions and 32 users filled that survey after going over the prototype. We used Firefox, Internet Explorer and Google Chrome along with other major browser to perform interface and compatibility testing. Performance testing was done by the online tool called Load Impact [24] for the homepage to analyze how many users can be handled by the server. Lastly we check for various server vulnerabilities to verify the system and the server is secure and has no bugs.

The survey asked the users to rate the navigation, understanding of the system, level of satisfation, concept of the website and overall system. In figure 9 and 10 x-axis shows the scale from 1-5 where. The y-axis shows the number of user's responses and they shows the navigation and overall system rating by users. Figure 11

shows the performance testing of the system. The y-axis is the delay in seconds caused in response of the server and x-axis is the simultaneous connections of the users. We can clearly see that system is responding fine and the delay is less than 1.5 seconds.

Figure 9 Rate Navigation

Figure 10 Rate Overall System

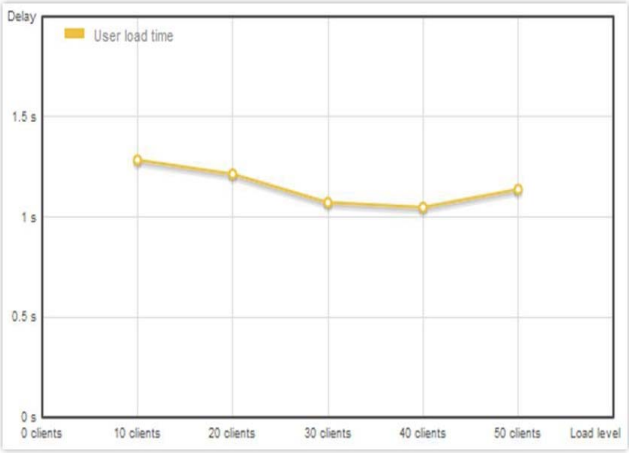

Figure 11 Performance Testing

Based on the analysis of the use test cases we can see that the prototype is functional and works properly. Users are rather satisfied with it and it had passed successfully the usability tests. The performance test was also fine as we can see that up till 40 clients the system was responding very quickly and for the 50 clients system responded but late. The more clients we would add will cause some delay but this delay is very minor. Currently we do not expect to have more than 1000 clients at the same time.

Although the system was generally accepted by the users and passed all the tests, we still have lots of room for improvements.

6. Conclusion and Future work

6.1 Limitations

Currently, there are some limitations but with the passage of time and efforts we would be able to overcome them. More precisely, we found two major restrictions:

- The Twitter API (Application Programming Interface [16]) [17] is not mature yet and its limiting the access of API per hour.
- Google does not provide any API for accessing its trend data also the hot trend feature of the Google generates hot trends but does not show the number of queries per hour.

6.2 Future Work

Our next phase of the prototype would include many additional features/improvements. Some of them are as follows:

- Integration of remaining search engines and social media websites like blogpuls [31].
- Semantic aspect of the search trends to be addressed.
- Normalization and scaling of data from different sources
- Visualization of the results improvements for better presentation of the data.
- Integration of Google maps and displaying trend on the basis of location.
- Decision support system to discover search trends patterns.
- Addition of value co-creation facilities.

6.3 Conclusion

The existing tools for the web search trends are mainly based on a particular search engine or social media website. They mainly provide search trend information pertaining to their own queries or discussions specific to their own platform. There is no existing tool that displays search trends from different search engines and social websites by combining them.

We have made an endeavor to build an ultimate market analysis tool which will facilitate the entrepreneurial activity of opportunity recognition through combination of web search trends from major social media websites and search engines.

ST tool as an integrated and independent platform shall provide the entrepreneurs with a broader spectrum of information over the web. Analysis of user interest through visualization of web search trends would be very useful in market analysis and decision making. As the architecture of the tool is flexible, more search engines and social media search trends according to their popularity and need can be added easily. ST tool has the scope for improvement by introducing semantics, decision support system and better visualization techniques.

7. REFERENCES

[1] Steve Lawrence, C. Lee Giles. Searching the World Wide Web Computer Science, NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA.

[2] Sergey Brin, Larry Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". Proceedings of the 7th international conference on World Wide Web (WWW). Brisbane, Australia. pp. 107–117.

[3] The Google Search Engine: Commercial Search Engine founded by the Originators of PageRank, http://www.google.com/, 2003

[4] Market Share (2010). Net Market Share SM Retrieved July 11, 2010, from http://marketshare.hitslink.com/

[5] Mika Kaki, Anne Aula, Controlling the complexity in comparing search user interfaces via user studies, Information Processing and Management 44 (2008) 82–91.

[6] Yair Shimshni, Erin Efron, Yossi Matias. On the Predictability of Search Trends, August 2009. Retrieved July 11, 2010, from http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/google_trends_predictability.pdf

[7] Ioannis Anagnostopoulos, A capture–recapture sampling standardization for improving Internet meta-search Computer Standards & Interfaces 32 (2010) 61–70

[8] J.P. Olhui. The social dimensions of entrepreneurship. Journal of Technovation, 2005 Vol. 25. Pp. 939–946

[9] Zheng Xiang, Ulrike Gretzel, Role of social media in online travel information search, Tourism Management 31 (2010) 179–188

[10] Sooper Tutorials (2010). Google SWOT analysis Retrieved July 11, 2010, from http://www.soopertutorials.com/business/strategic-management/1369-Google-swot-analysis.html

[11] Marketing Teacher (2010). Yahoo SWOT analysis Retrieved July 11, 2010, from http://marketingteacher.com/SWOT/yahoo_swot.htm

[12] Vishal Jhaveri. Social Media Marketing Strategy, Youth Noise (2010). Retrieved July 11, 2010, from http://youthnoise.com/karoline/YNstrategy_SAMPLE.pdf

[13] Briansolis (2009). Facebook Top trends of 2009 Retrieved July 11, 2010, from http://www.briansolis.com/2009/12/Facebook-top-trends-of-2009/

[14] Twitter (2010). Twitter 2009 trends Retrieved July 11, 2010, from http://blog.Twitter.com/2009/12/top-Twitter-trends-of-2009.html

[15] Zend (2010). Zend Studio Retrieved July 16, 2010, from http://www.zend.com/products/studio/

[16] PC Magazine (1996). Definition of: API Retrieved July 16, 2010, from http://www.pcmag.com/encyclopedia_term/0,2542,t=application+programming+interface&i=37856,00.asp

[17] Twitter (2010). The Twitter API Retrieved July 16, 2010, from http://apiwiki.Twitter.com/API-Overview

[18] Trendistic (2010). Seeing trends Retrieved July 16, 2010, from http://trendistic.com/_help

[19] Systems Engineering Fundamentals. Defense Acquisition University Press, 2001

[20] SimpleTest (2010). Simple Test Retrieved July 16, 2010, from http://www.simpletest.org/

[21] Sourceforge (2010). PHP Unit Retrieved July 13, 2010, from http://phpunit.sourceforge.net/

[22] Last Craft (2010). Simple Test for PHP Retrieved July 16, 2010, from http://www.lastcraft.com/simple_test.php

[23] Abdul Wahid, Asma Rafiq, Farooq Ahmad, Petko Ruskov. "Discovering Business Opportunities via Search Trends" May 27, 2010, International Conference Entrepreneurship, Innovation and Regional Development, pp. 818-826.

[24] Load Impact (2010). Introduction Retrieved July 16, 2010, from http://loadimpact.com/info/introduction.php

[25] Netcraft Ltd (2010). June 2010 Web Server Survey Retrieved July 16, 2010, from http://news.netcraft.com/archives/2010/06/16/june-2010-web-server-survey.html

[26] Search Engine Watch (2010). What people search for Retrieved July 16, 2010, from http://searchenginewatch.com/2156041

[27] Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007.

[28] Google (2010). About Google Trends Retrieved July 16, 2010, from http://www.google.com/intl/en/trends/about.html

[29] Google (2010). Google insight Retrieved July 16, 2010, from http://www.google.com/insights/search/

[30] Google (2010). Google Trends Retrieved July 16, 2010, from http://www.google.com/intl/en/trends

[31] Blogpuls, available at: http://www.blogpuls.com [Cited 16th July 2010]

[32] Choi, H., & Varian, H. (2009). Predicting the Present with Google Trends. Retrieved from http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf