



GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY
Volume 11 Issue 20 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Does a Citation-Index for Websites Make Sense?

By Martin Ebner, Aldi Alimucaj

Graz University of Technology

Abstract - The usefulness of external references to websites is a crucial factor of web-monitoring. It is of high general interest whether your website is visited by the estimated target group or not. This publication measures the value of references of websites by bringing the idea of citation-index to web-analytic tools. The approach presented is considering the number and quality of actions a visitor of a specific website does and the time s/he spent on this website as well as the previous website the user is coming from. The combination of these three parameters are expressed by formulas and afterwards visualized by different tools. Finally this approach is tested and discussed on an existing project. It can be concluded that this concept is indeed useful to get a deeper insight whether external websites addressing the intended target group or not.

Keywords : citation-index, ranking websites, web analytics, L3T, piwik.

GJCST Classification : H.3.1, H.3.5



Strictly as per the compliance and regulations of:



Does a Citation-Index for Websites Make Sense?

Martin Ebner^α, Aldi Alimucaj^Ω

Abstract - The usefulness of external references to websites is a crucial factor of web-monitoring. It is of high general interest whether your website is visited by the estimated target group or not. This publication measures the value of references of websites by bringing the idea of citation-index to web-analytic tools. The approach presented is considering the number and quality of actions a visitor of a specific website does and the time s/he spent on this website as well as the previous website the user is coming from. The combination of these three parameters are expressed by formulas and afterwards visualized by different tools. Finally this approach is tested and discussed on an existing project. It can be concluded that this concept is indeed useful to get a deeper insight whether external websites addressing the intended target group or not.

Keywords : citation-index, ranking websites, web analytics, L3T, piwik.

I. INTRODUCTION

Ranking information is one of the most important issues on the web, bearing in mind that there is no good or bad information. By considering this and by knowing some conventional methods such as the *impact factor* for ranking scientific journals, following research question centered our interest: *does a citation-index for websites make sense?* The citation-index works similarly to existing ones but with a special focus on your publication. In other words it does not compare the ranking of your publication to other publications directly it relatively rank how often publications got cited by others. Assigned the same principle for websites a better understanding of referencing websites sharing same interests will occur and deepen the cooperation between websites in order to reach a win-win situation. Unlike the impact factor our citation-index is conceived for the web, so when we speak about other sites we mean websites like ours, which quotes us and adds the link as a reference. This link as we will explain later on is the key for measuring the criteria we set up for the index. Beside the theoretical part we took advantage of the online book called L3T (German textbook about Technology Enhanced Learning) (<http://l3t.eu/>) to gather the necessary information for making a founded reasoning about the advantages this new index offers.

Author^α : Associate Professor at the Institute of Information Systems Computer Media and head of the department Social Learning at Graz University of Technology. E-mail : martin.ebner@tugraz.at

Author^Ω : bachelor degree in Computer Science and Economics and is currently finishing his master degree at the University of Technology in Graz. Telephone: 00436802305474
E-mail : aldi.alimucaj@student.tugraz.at

II. THEORY

a) Impact Factor

Publishing in scientific journals is very important for the career of a scientist. Choosing the right journal may be crucial for that. Of course there are many journals and they for sure differ in quality which is hard to evaluate. But easy enough, to be found just by doing some bibliographic research and counting the number of citations of articles published in a specific journal. One tool for estimating the relative prestige of journals in a given field is called Journal Citation Reports. JCR is an electronic resource which determines the frequency of citation in total, average as well as the impact factor. The impact factor of a journal is among the criteria considered when candidates are evaluated for promotion [Day, Gastel 2011 p. 30].

DEFINITION

"The impact factor is a measure of the frequency with which the "average article" in a journal has been cited in a particular year or period. The annual JCR impact factor is a ratio between citations and recent citable items published. Thus, the impact factor of a journal is calculated by dividing the number of current year citations to the source items published in that journal during the previous two years (see Figure 1)."

A = total cites in 1992

B = 1992 cites to articles published in 1990-91 (this is a subset of A)

C = number of articles published in 1990-91

D = B/C = 1992 impact factor

Figure 1 : Calculation for journal impact factor (Source: thomsonreuters.com, July 2011)

However this index has its limitations as well, for example it reflects just the impact factor of the whole journal not of individual articles. It is not interdisciplinary and cannot measure journals of different fields. It is obvious that some journals get a higher rating by counting replies to articles that cite the article in question but not counting them as papers. Editors can increase the impact factors of their journals by publishing good polemical articles early in the year [Hartely 2008, p137].

b) Our approach

First the general rules must be set up and combined them together into an equation to form a ranking system. Many of the web analytics systems listed below such as *google analytics*, *piwik* or *open web analytics* are offering all the data which can be tracked from the user (called "raw data") but preview them in no relationship with each other. This was the purpose of our study, to build a system, gather data, analyze them and give conclusions about the possibility of its application. First the web analytics framework is introduced which helps us gathering the necessary data together with a brief introduction to web analytics itself.

c) WEB ANALYTICS

The Web Analytics Association (<http://www.webanalyticsassociation.org>) has proposed a standard definition for web analytics:

"Web analytics is the objective tracking, collection, measurement, reporting, and analysis of quantitative Internet data to optimize websites and web marketing initiatives." [Kaushik, 2007 p. 6]

Following this definition, collecting data is just one of many functions web analytics can and has to fulfill. The data that are being collected and measured are called clickstream information. Clickstream is foundational data that helps to measure and analyze all kinds of site behavior: visits, visitors, dwell time on site, page views, bounce rate, sources, and more. On base of these data we can analyze the following aspects:

- Brand buzz and opinion tracking
- Customer satisfaction
- Net promoter indices
- Open-ended voice-of-customer analysis
- Visitor engagement
- Stickiness
- Blog-pulse

There are many business models that use web analytics for their selling and/or promoting purposes. Whether it is an online shop, a blog or some highly specialized financial software that runs on the browser.

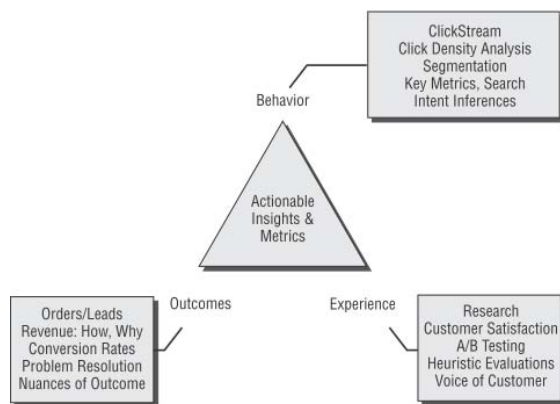


Figure 2 shows the trinity framework; a new way of perceiving web analytics for most efficient data outcome. The goal of behavior analysis is to infer the intent of the website visitors basing on all what we know about them, which is basically clickstream data. The outcome is the result measured in company's predefined objectives, for example if it is an e-commerce website, how many viewers did actually buy the product. But for detailed information analysis and understanding customers' behavior we need profiled web analytics.

i. Web Analytics Frameworks

Web analytics tools date back in the early 90s. Since then they have been improving from simple requests counting to highly accurate JavaScript clients, from log files to databases and from text outputs to impressive reporting methods. Besides commercial tools (see Table 1) there are some well implemented open source competitors as well.

COMMERCIAL TOOLS FOR WEB ANALYTICS

ClickTracks	ClickTracks provides an innovative line of products and hosted services in the field of web site traffic analysis. ClickTracks focuses on presenting meaningful information about user behavior visually in context.
Coremetrics	Coremetrics Web Analytics platform captures and stores all customer and visitor clickstream activity to build LIVE (Lifetime Individual Visitor Experience) profiles that serve as basis for all successful e-business initiatives.
Google Analytics	Google Analytics offers free web analytics services with integrated analysis of Ad-Words and other keyword-based search advertising. Google Analytics bases on Urchin, which Google purchased in 2005.
NedStat	NedStat is a provider of software solutions and services for monitoring websites and reporting on website-visits.
Omniture	SiteCatalyst is a hosted application that offers a comprehensive view of activity on a company's website that includes historical (data warehouse) and real-time analysis as well as reporting. SAS Web Analytics applies SAS Customer Intelligence software to online channels for a complete view on the customer's interaction.

Figure 2 : The trinity diagram. Source: [Kaushik, 2007, p. 18.]

Visual Sciences	Real-Time Analysis Platform (RTAP) and Suite of applications for to collect, process, analyze and visualize user data for decision making; including Internet sites and services.
WebTrends	WebTrends offers both an on demand service as well as software solutions for measuring campaign performance, search engine marketing, web site conversion and customer retention.

Table 1: Commercial Web Analytics Frameworks.
[Source: digitalenterprise.org, July 2011]

Two of the most popular open source web analytics tools are Piwik (<http://piwik.org/>) and Open Web Analytics (<http://www.openwebanalytics.com/>). They are both licensed under GPL (www.gnu.org/copyleft/gpl.html) and offer nearly the same features and use the same technologies. We implemented our plugin for Piwik which is the framework we are going to discuss in details.

a) *PIWIK*

Piwik is a downloadable, open source (GPL licensed) web analytics software program. As an alternative to services like Google Analytics, Piwik allows you to host your statistics services on your own server, have full ownership and control over the data collected from your visitors. A plugin offers a user interface which is very manageable and easy to use.

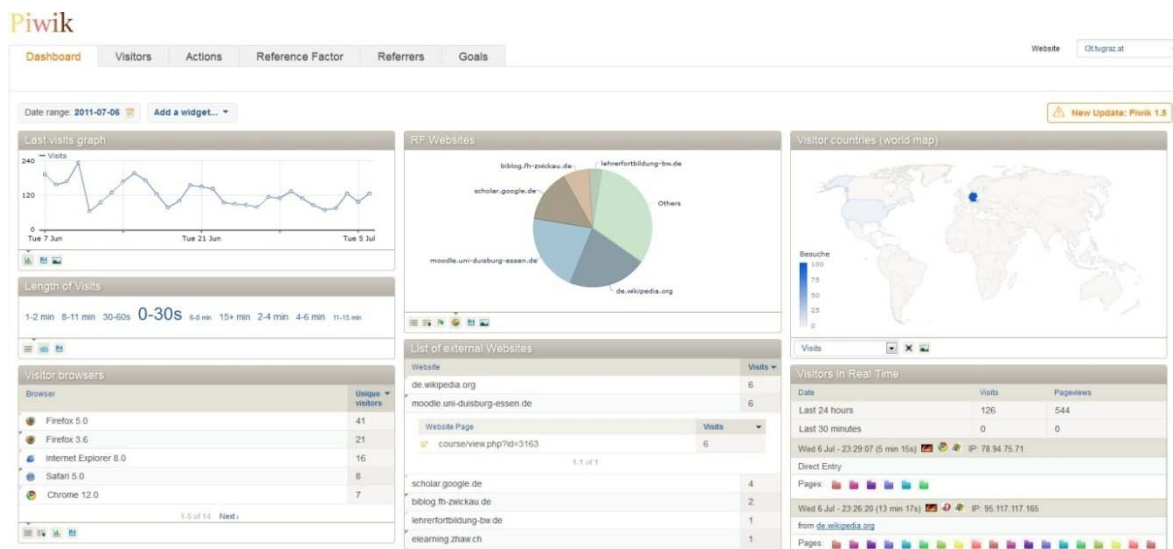


Figure 1 : Piwik user interface. [Source: l3t.tugraz.at Piwik]

Piwik is currently on stable version 1.5 and offers many features such as:

- o Real time reports
- o Detailed view of your visitors, pages they have visited, goals triggered
- o Customizable dashboard
- o Dashboard for all websites
- o Analytics for ecommerce
- o Ecommerce – abandoned carts reporting
- o Ecommerce – product and categories reporting
- o World map of your visitors
- o Automatic tracking of file downloads
- o Automatic tracking of clicks on external websites
- o Analytics campaign tracking
- o More than 800 search engines tracked
- o Scheduled email reports (PDF and HTML reports)

[Source: piwik.org, July 2011]

b) *EXTENDING PIWIK*

Besides the user interface Piwik offers a lot of plugins out of the box which cover most of the customer wishes. One of the tasks of this work was to implement our own metrics and incorporate them into Piwik through its plugin interface. First thing to do was to understand Piwik's architecture, how could it be extended and what possibilities does it offer.

c) *TECHNICAL IMPLEMENTATION*

Piwik gathers all its information from a JavaScript client called the Tracker Code which is anchored into the websites that need to be observed. When a user opens the site it sends the initial information to the server containing browser specification, OS platform, language, forwarding link and so on. After that, the client continues polling information about user activities such as click actions or time spent doing something. On the server side Piwik

has a well-developed MVC (Model-view-controller) architecture based on Zend Framework (a PHP optimizing package). The plugins are on the other hand based on the MVC architecture themselves and can be seen as application within the application. From the database we accessed most of the data needed for our new metrics so we didn't need to collect new data from the client.

d) DEFINING THE METRICS

The Tracker Code supplies the forwarding external website's name which linked the user to the website of our interest. The main idea is to find out those external websites that forward the most fitting target group regarding to the website analyzed. In other words, which external website should we set our focus on and is worth invested more time on? There are different parameters how to measure that: First to find out is how many users are coming from a specific website to ours, second how many actions does the user do on our website and third how long did the user stay on our website. The goal of this research work is to combine these three parameters into one formula and visualize the usefulness of references pointing to our website.

In general Piwik offers the raw data to build a more complex analysis. The first formula to be applied was intended to show the average values of incoming connections for a given time frame. Since we are working on the incoming references from other website we called it "Reference Factor" (RF). The average reference factor formula is shown in Formula 1

$$RFA(w) = \frac{\emptyset V_w}{\emptyset V_a} * \frac{\emptyset A_w}{\emptyset A_a} * 10^6$$

Formula 1 : Reference Factor Average

Furthermore a second formula called the "Multiplicative Reference Factor" (RFM) is needed, because we first define the ratio of the website data with the system and then multiply the data in order to set them in relation with each other.

$$RFA(w) = \frac{V_w}{V_a} * \frac{A_w}{A_a} * 10^2$$

Formula 2 : Reference Factor Multiplicative

The given shortcuts are explained in Table 2.

RFA	Reference Factor Average
RFM	Reference Factor Multiplicative
V	Visit Time
A	Actions
B _w	Visits website
B _a	Visits system
∅	Average
V _w	Visit Time website
V _a	Visit Time system
A _w	Actions website
A _a	Actions system
W	Website

Table 2 : Formula abbreviations

The average RF is build up with the average values of the reference website (rw) and those from the system. To build an average value it is important to create a rank, which is based on quality instead of quantity. For example we know that website1 is at the first place and has for example 447 visits over 6 months with 2045 actions and 94789 seconds visit time. But measured in average values website2 with 39 visits and 474 actions 42108 visit time has much more interested users who are willing to spend more time on our website. This tells us that website1 users could be misled or were just lurking but website2 users where certain of the content and found just what they were looking for. The RFM is a measurement scale involving visit time and actions, brought together to build a benchmark. A hugh number of users coming from a website is leasing to a high RF-factor of that site, so all popular sites are always at the top. That's why we have to consider both diagrams for an accurate overview. Both formulas were multiplied with factors of 10 to improve their conspicuity.

e) Reporting mechanism

RF Websites					
Website	Visits	Actions	Visit Time	RF Multipl	RF Avarage
de.wikipedia.org	460	1913	75795	108	52
www.facebook.com	447	2045	94789	144	74
scholar.google.de	431	1440	69138	74	41
www.weiterbildungsblog.de	339	1459	71861	78	70
twitter.com	287	1035	41093	32	40
www.futurezone.at	194	876	38044	25	68
konzeptblog.joachim-wedekind.de	192	477	19879	7	20
www.medienpaedagogik-praxis.de	188	956	44313	31	92
blog.studiumdigitale.uni-frankfurt.de	178	653	30549	15	48
www.tu-chemnitz.de	155	853	27806	18	76

1-10 of 444 Next

Figure 2 : Piwik RF Widget

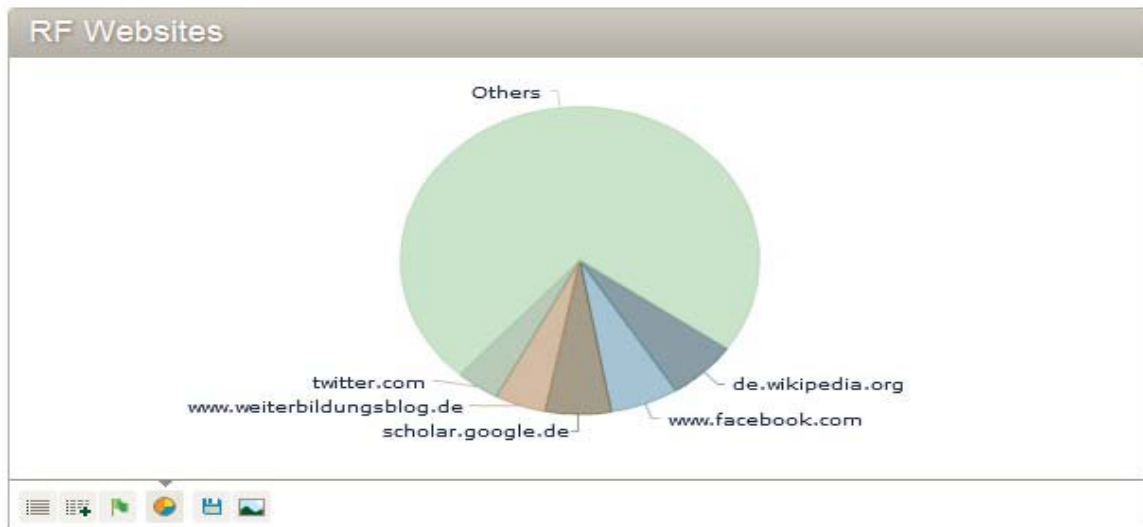


Figure 3 : Widget, Pie Chart View

Figure 3 shows the implemented widget. It builds a table of reference websites and their corresponding number of visits in the first column, the actions, visit time, RFM and RFA. The table can be browsed and sorted. You can even “unfold” one site and take a look at the link where your tracker is placed. You can even build pie charts or vertical bar graphs within

the widget. These graphs were satisfactory for one dimensional values. But the RF-s where compound of many values, so it had to be multi-dimensional. To fulfill that we chose a powerful tool (such as MS Excel) to build the graphs out of three dimensions: actions, visit time and visits. The yearly graph looks like figure 5 and 6

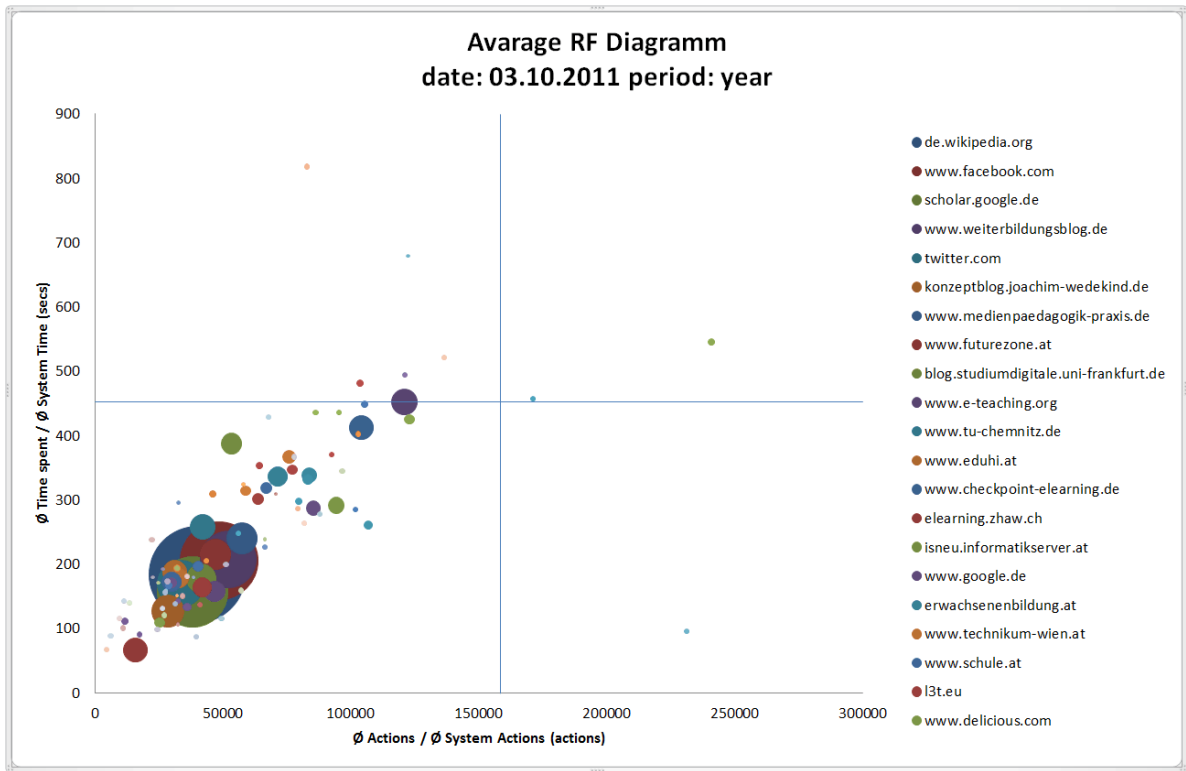


Figure 4 : RFA period: year 2011

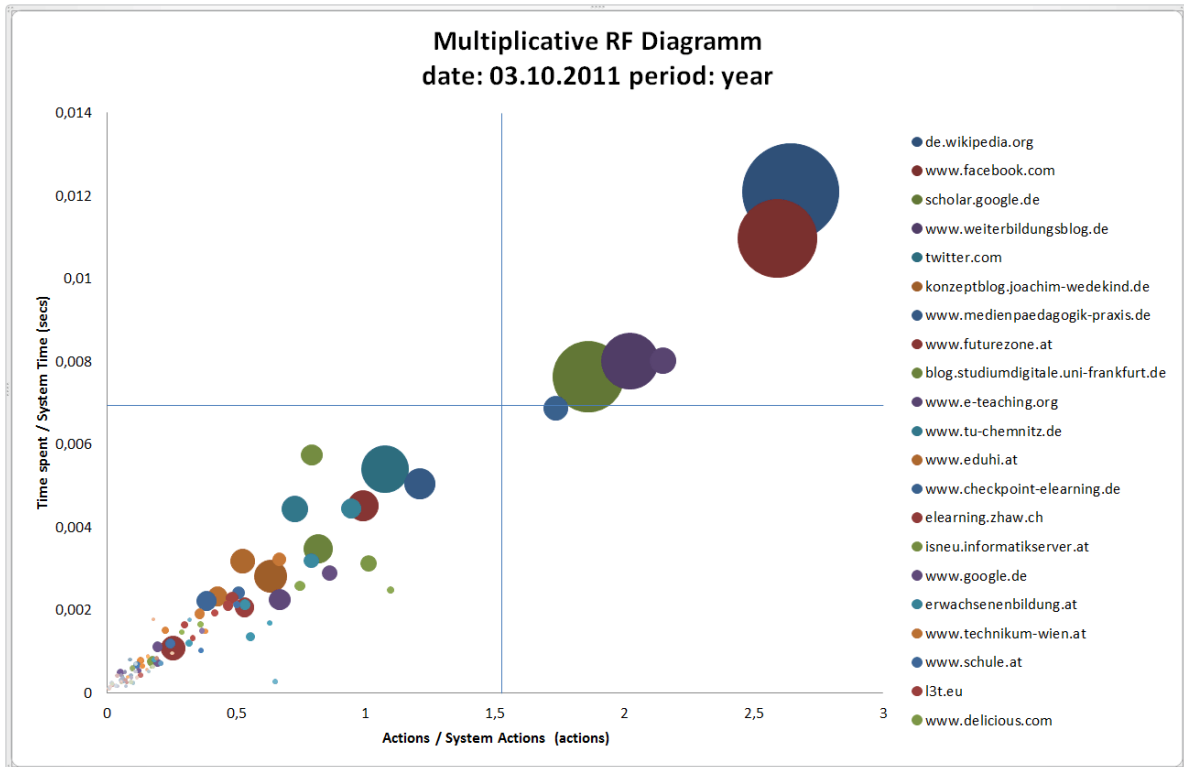


Figure 5 : RFM period: year 2011

Microsoft Excel didn't offer the possibility to build these three dimensional graphs out of the box where the third one is the number of visits represented by the size of the circle. So the only possibility was to write a VBA Macro (Visual Basic for Applications) to build the graph and make the changes we needed. Some items are underrepresented with a smaller circle size as it can be shown so we set a default representation value for all items smaller than three units. The size of the circle is also adapted to the graphs possibilities like i.e. some site with more the 1000 visits cannot have a circle of size 1000 because it would be too big to be rendered. So the biggest value is divided by the scale of max circle size and the rest is adapted to that value. The Macro is tested in MS Excel 2011 and MS Excel 2010 for Mac and Windows.

III. PROOF OF CONCEPT

Having finished the technical implementation we tested the whole concept and prove its capabilities. The project L3T (<http://l3t.eu>), which is a German text book on technology enhanced teaching and learning, was chosen for that purpose. The project website of L3T already had piwik installed providing enough data for a good analysis and conclusions. In piwik 1.1 the possibility is given to choose between fixed time periods which are daily, weekly, monthly, or the whole year. Of course, the more data you have to process the more accurate becomes the result. So we started with the annual period which rendered us the table as to be seen in figure 3. On base of this table we can get all the information needed to work out a conclusion. If the same information is needed for example for a presentation then the content could be exported and rendered in MS Excel with the supplied macro. The data used for this table is the same as for the simple ranking. The application of the reference factors offers us two new types of ranking which could be similar but don't have to be the same with the simple ranking.

IV. DISCUSSION

As mentioned before, the tables give an accurate data but lack on fast visual interpretation potential; so for the final discussion the excel graphs will be used. Taking a look at the monthly periods we see that those are more dynamic and could reveal information that is smoothened by larger time periods.

February 2011 was the first monthly information gathering period to start. It can be shown that popular websites like *facebook* very soon attracted a large number of users. But it didn't take long for other sites to contribute to the popularity of the project. In fact for the rest of other monthly reports other profiled websites such as those from universities or Wikipedia were running on top of the RFM list. This is because users who were linked through those websites where more interested in the topic which resulted in longer visit times

and more actions within the website. Browsing and long time reading means that the user found what s/he was searching and looking for. In March 2011 for example many other websites were represented as big contributors by reaching the right-top side of the RFM graph having larger circles. Although they generated about the same amount of visits during the time, the quality (actions, visit time) were not always predictable. Wikipedia accords to the yearly period graph at first place in raw data measurement and RFM because it generated a lot of traffic. And a lot of traffic means many users have visited the site and are familiar with its content. They might not have found what they needed but they know what L3T is about and would take a reference on it the next time they would need it. The average RF at the other hand tells us about the interest of the user despite the number of visits. When we look at figure 7 displaying the RFA of June we can see a small dot at the top-right edge of the graph. This dot represents moodle.uni-graz.at at the first place for the monthly ranking. Although *facebook* with a larger circle has the highest number of visits on average it has a smaller RFA quotient than *moodle*. Moodle only forwarded one visitor that month but that person was so interested in the page that s/he spent over 40 minutes reading taking over 20 actions, which are far up from *facebook's* average values.

So finally, it must be pointed out that the best way to tell the importance of a site is, if it ranks in the same area in both diagrams. This is for example the case for www.checkpoint-elearning.de (154 visits, 1330 actions) and www.e-teaching.org (164 visits, 1554 actions) in the yearly diagrams (Figure 5 and Figure 6). They have more or less the same amount of visits and actions with a similar ratio between both. For such sites we can draw the conclusion that quantitative and qualitative values are valid. It can also be stated that if one site is positioned in the upper right area of the diagram than it offers interesting potential in our sense. So for sites occupying the same area we can estimate that the assumption we intended with our research question is true, but for the rest of sites the fluctuation is too big to make a clear distinction.

Besides the monthly reports there are weekly analysis as well. The amount of data is relatively small for drawing conclusions but is sufficiently meaningful for staying up to date with the newest developments regarding your website's popularity. It can also be used for history purposes to compare relevant changes.



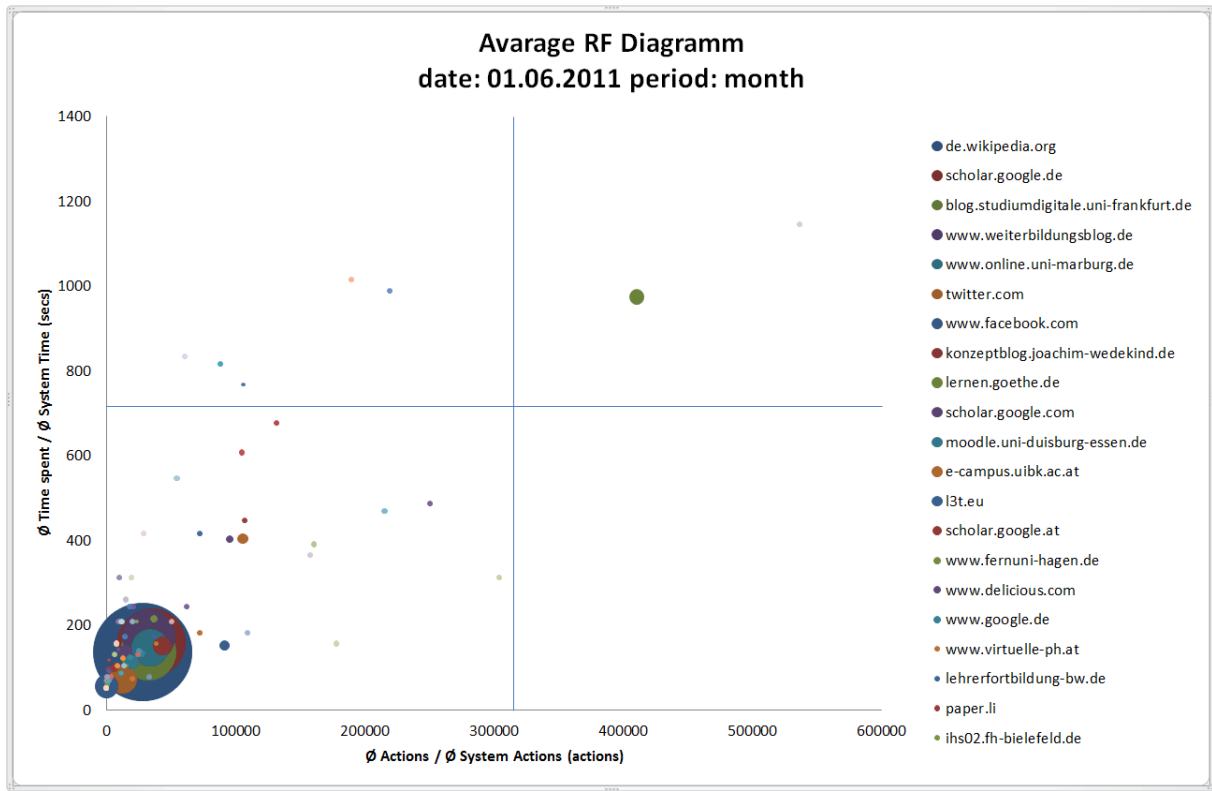


Figure 6 : RFA period: June 2011, month

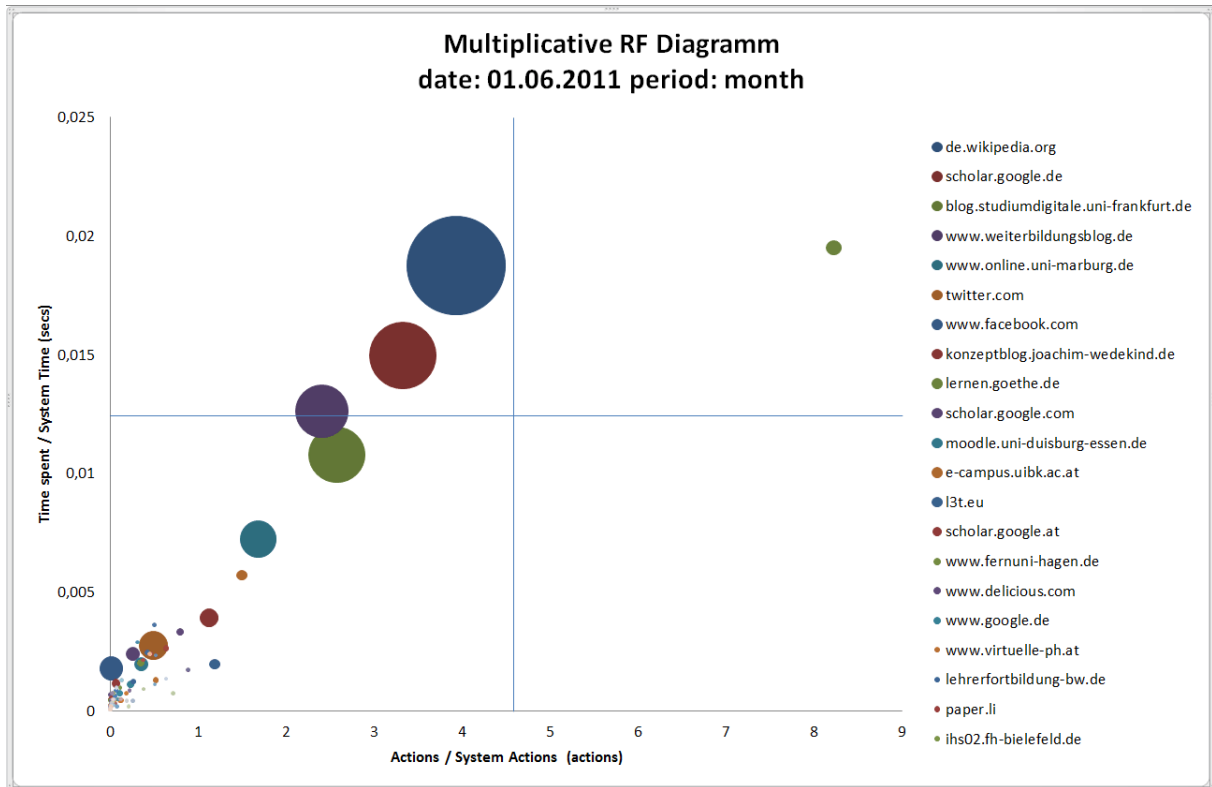


Figure 7 : RFM period: June 2011, month

V. CONCLUSIONS

This paper focuses the question whether web analytic tools help us to filter relevant web visitors by interpreting their link history. Therefore two new measurement methods to rank the effectiveness of reference websites were set. Furthermore these methods were implemented into an analytics system and by exporting the data complex graphs could be built with the help of external tools.

On base of such reports two different factors were calculated. The first one was the multiplicative reference factor which results from bringing raw data in connection with each other and the second one was the average reference factor as an outcome of the average values. The tested example (L3T project) has shown that the final diagrams help to interpret the usefulness of external references to the example project website. Web analytics remain a big field for online-business. Ranking systems will become more sophisticated trying to differentiate real chances to separate from the noise of Internet.

REFERENCES REFERENCES REFERENCIAS

1. A Day, R. (2011) How to Write and Publish a Scientific Paper, Greenwood ABC-CLIE LLC
2. Ebner, M., Schön, S. (2011) *Lehrbuch für Lernen und Lehren mit Technologien*, BookOnDemand, Germany, <http://l3t.tugraz.at>
3. Hartley, J. (2008) *Academic writing and publishing: a practical guide*, New York, Routledge
4. Kaushik, A. (2007) *Web Analytics*, Wiley Publishing, Inc., Indianapolis, Indiana

HYPERLINKS

digitalenterprise.org, Managing the digital enterprise, 12.07.2011 [online] <<http://digitalenterprise.org/metrics/metrics.html>>

thomsonreuters.com, The Thomson Reuters Impact Factor, 12.07.2011 [online]

<http://thomsonreuters.com/products_services/science/free/essays/impact_factor/>

piwik.org, Piwik Open Source Web Analytics, 12.07.2011 [online] < <http://piwik.org/features/>>

