Mobile Web Profiling: A Study of Off-Portal Surfing Habits of Mobile Users

Daniel Olmedilla, Enrique Frías-Martínez, and Rubén Lara

Telefónica Research & Development
Emilio Vargas 6, 28043
Madrid, Spain
{danieloc,efm,rubenlh}@tid.es

Abstract. The World Wide Web has provided users with the opportunity to access from any computer the largest set of information ever existing. Researchers have analyzed how such users surf the Web, and such analysis has been used to improve existing services (e.g., by means of data mining and personalization techniques) as well as the generation of new ones (e.g., online targeted advertisement). In recent years, a new trend has developed by which users do not need a computer to access the Web. Instead, the low prices of mobile data connections allow them to access it anywhere anytime. Some studies analyze how users access the Web on their handsets, but these studies use only navigation logs from a specific portal. Therefore, very little attention (due to the complexity of obtaining the data) has been given to how users surf the Web (off-portal) from their mobiles and how that information could be used to build user profiles. This paper analyzes full navigation logs of a large set of mobile users in a developed country, providing useful information about the way those users access the Web. Additionally, it explores how navigation logs can be categorized, and thus users interest can be modeled, by using online sources of information such as Web directories and social tagging systems.

1 Introduction

Nowadays, millions of users in the world access daily information on the World Wide Web. We have transitioned from a Web available only to the academic world to a large source of data available to almost everyone. This transition has produced a new range of services combined with new business models (e.g., online advertising). A significant amount of effort has been dedicated to analyze and classify the activity users perform on given web sites (on-portal) or within their whole navigation sessions (off-portal) in order to profile and improve their experience (e.g., through personalization) and the web site service quality (e.g., sponsored links, recommendations, targeted vs. non-targeted advertisement). This work includes generic studies of surfing behavior and patterns [1–3], and more specific ones such as extracting profiles from Web navigation logs, link analysis for personalized Web Search or recommendations [4–6].

P. De Bra, A. Kobsa, and D. Chin (Eds.): UMAP 2010, LNCS 6075, pp. 339–350, 2010.

Lately, mobile phones have become a part of our daily life (in fact there are around 4 billion users subscriptions in the world[1] and around a billion new phones are bought each year[2], including new subscriptions and handset replacements). With new 3G technology and the reduction of data connection tariffs, not only users can access the Web from their handsets, but they are doing it constantly and in new set of situations not possible before (on-the-go). However, due to the lack of available data, only work with on-portal (logs generated within a portal) on user web navigation analysis and profiling has been performed. Very little work has been performed in order to analyze the usage users make of their mobile Web navigation capabilities in a broader spectrum, that is, analyzing their navigation from the moment they connect to Internet to the point where they disconnect, independently of how many portals they have visited.

This paper shows a first effort on trying to understand users off-portal mobile navigation behavior. We analyzed the logs of three months of web navigation (52 million visits to Web domains by 283,000 users) in a developed country and identified, among others, which type of web sites users accessed, which distribution these visits followed and which main categories users are interested in. This information is very valuable in order to personalize services (e.g., better knowledge of the customers in order to improve on-line recommendations and advertisements) as well as to improve existing ones (e.g., parental control services which filter out or warn users when accessing sites with adult or inappropriate content). However, extracting the main categories a domain belongs to is not an easy task.

In this paper, instead of text mining web pages, we tried to use collective intelligence available on the Web in order to automatically classify web sites. First, we relied on the Open Directory Project[3], which provides the largest manually annotated Web directory, and classifies web pages within a total of 17 top categories. Additionally, we also accessed information available on social tagging systems such as YahooMyWeb[4], as well as existing meta-tags available in the accessed web pages. Combining this two sources of information (categories and tags) allows us to identify a set of representative tags per category, which can be used to classify new domains for which tags exist, but no manual categorization has been performed.

The rest of the paper is organized as follows: Sect. 2 presents previous related work in the area. The dataset used and the first analysis made over it are presented in Sect. 3. Section 4 introduces a categorization scheme of pages found in the navigation logs based on the Open Directory Project, and analyzes its results. Web page Meta-Tags and social tagging systems are exploited in Sect. 5

[1] http://www.ngrguardiannews.com/compulife/article02/141009, http://www.cellular-news.com/story/printer/32073.php

[2] http://communities-dominate.blogs.com/brands/2009/02/bigger-than-tv-bigger-than-the-internet-understand-mobile-of-4-billion-users.html

[3] http://www.dmoz.org/

[4] Information of YahooMyWeb was obtained before its shut down in March 18th, 2009.

in order to characterize pages, and Sect. 6 combines category and tag information in order to infer new categories from pages that otherwise would not be categorized. Finally, Sect. 7 concludes the paper and outlines future ideas we plan to explore in the future.

2 Related Work

The WWW has already been studied and characterized for online access in a variety of studies [1–3]. However, little is known about the behavior of users when navigating the Web through their mobile phones.

The literature reports a variety of studies that characterize user navigation, search strategies and content for mobile Internet. One of the main characteristics of such studies is that the data used for the studies typically comes from just one portal. In this context, [7] shows that the law of web surfing [3] (developed for traditional on-line access) holds true also for mobile web access, using an extensive data set coming from one web site. The studies presented in [8, 9] show the dynamics of mobile access to a commercial web portal, finding, as previous studies did, that the majority of client request are for a reduced number of documents (i.e. the navigation patterns are very similar for all mobile users). The work presented in [10] focuses on studying search patterns in Google for mobile users. Their conclusions indicate that the diversity of queries in mobile access is far less than in desktop, and that although users for the best part search similar content in both environments, the percentage of Adult queries is vastly larger in mobile access. Based on these results a variety of applications have been developed for predicting user navigation[11] and adapting content for mobile users [12].

A characteristic of mobile access characterization studies is that, while desktop access can be considered homogeneous, mobile access is done with phones with different capabilities (ranging from the size of the screen to the data input method) that deeply affects the analysis. For example, [13] found out that although searches in mobile phones are much shorter than in computers, searches done from iPhones were very similar to the ones performed through a computer, being this conclusion also true when evaluating the variety of queries.

There are some studies that use more than one portal to characterize the mobile WWW, nevertheless those studies focus on characterizing content. For instance, [14] studied over one million mobile pages, and found that from the three content types (WML, C-HTML and XHTML) WML was dominant.

To the best of our knowledge our study is the first one that characterizes and studies mobile user navigation using navigational data originating from the user not from a portal. This implies that for each given user our analysis reflects the different navigation sessions over the portals that the user has accessed for the period of time considered. In this context our study complements previous results in analyzing and characterizing user behavior that have focused (due to the complexity of obtaining the data) on an individual portal.

3 Mobile Web Navigation Analysis

As the basis for the analysis that is described in this paper, information from users off-portal access to mobile Internet via handsets has been used. This dataset includes a total of 52 million accesses (visits hereafter) to more than 45,000 different domains by more than 283,000 users for a time period of 3 months. This usage data belongs to a developed country and the information used from these logs include for each entry an anonymised user id, the domain accessed (not the whole URL in order to preserve privacy) and the time when the access took place.

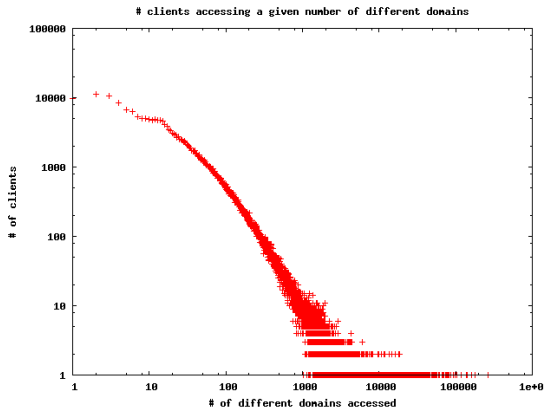

Fig. 1. Distribution of number of domain accesses by clients

Figure 1 and Table 1 respectively shows the distribution of domains visited by users and describes some basic statistics in order to better understand the nature of this dataset. In Fig. 1 we can observe a linear relation in a logarithmic scale between the number of users and the number of domain accessed, representing in a linear scale the typical long tale behavior, i.e. there is a core set of domains heavily accessed. This general behavior is also true when accessing the internet from laptops, and is in agreement with the results presented in [7].

The domains listed in the mobile web logs include a large number of domains that are only accessible from the handset, that is, via the operator network and not from a regular computer Internet connection. These domains typically represent portals belonging to the operator itself (or related advertising banners) or versions of websites adapted to mobile access. For the rest of the study, these domains are ignored, since there will not be any possibility to link them with the category and social tagging sources of information available on the Web. Therefore, we also eliminate the data associated to the users that only accessed those domains we are leaving out of the study, which amounts to 4.49% of our dataset. These customers are likely to have connected by mistake or in order to just test the connectivity, so they reached only the operator portal shown as

Table 1. Generic statistics associated to the analyzed mobile web navigation logs

Description	Amount	% over total
Total Users	283,198	
Total Accesses (visits)	52,297,157	
Total Domains Visited	45,103	
Web accessible domains	34,791	77.14%
Only-mobile domains	10,312	22.86%
Users visiting Web accessible domains	189,283	66.84%
Users visiting only-mobile domains	273,311	96.51%
Users visiting only-mobile domains	93,915	4.49%

starting page (also suggested by the fact that 96.51% of the users have visited at some point one of such operator domains and 90.14% have visited the operator starting portal).

4 Categorizing Web Domains

When analyzing user behavior within a portal, the categorization is typically performed by first categorizing the web pages (or areas) of the portal and then classifying each user visit according to the category assigned to that page. This can be done within a portal since the owner of the portal knows and has control over the content displayed on it. In these cases, user profiles can easily be constructed as a set of:

- (a) categories the user is interested in and
- (b) a weight for each category, typically based on the frequency pages on each category are accessed as a function of time (so recent visits are considered more important than older ones)

However, off-portal implies that users will access pages that are (most likely) not known, and therefore not classified, in advance. In these situations, user profiles typically consist of a set of top keywords extracted from the pages she has visited, and sometimes, there is an effort to classify those keywords in a set of categories. In this paper, we have decided to follow the opposite process. We will first try to analyze those webpages for which we know the category they belong to, and later on try to analyst which keywords are typically representing such categories.

In order to classify web pages on the World Wide Web, we relied on the Open Directory Project (ODP), "the most comprehensive human-reviewed directory of the web", which contains manual classification (by more than 85,000 editors) of more than 4 million sites over 590,000 categories (organized as a tree with 16 top categories). Therefore, we searched within the ODP database for the domains users were navigating. Table 2 describes the results of this analysis. As it is presented, only 15% of domains are annotated with at least one category within the ODP directory.

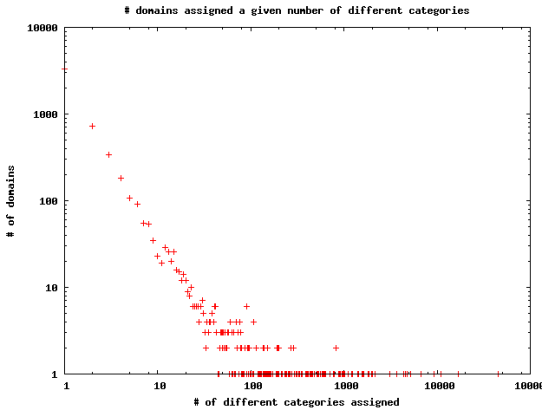

Fig. 2. Distribution of number of domains with N categories assigned

Table 2. Categories matches for domains and users accessing those domains

Description	Amount	% over total
Different Categories assigned to a domain	283,198	
Total Domains with a category assigned	5,483	15.76%
Total Users accessing a domain with category	37,826	19.98%

Additionally, since a domain might fall into more than one category, we analyzed how this distribution look like. Figure 2 shows the distribution of domains that are assigned a given number of categories, which as we can observe, follows a power-law distribution, i.e. a core of domains fall under a great number of categories (capturing probably news portals and such), while the best part of domains are assigned a reduced number of categories indicating also that the content of the portal is more focussed.

5 Page Meta-tags and Social Tagging Systems

In the past, the Web 1.0 provided a small number of people with the ability to share information with the whole world being able to access it. Website administrators (or some users via appropriate content management systems) could upload content to the Web and make it publicly available. Nowadays, the boom of the Web 2.0 allows users to act not only as consumers of information but also as providers. Wikis, blogs, community sites or photo sharing sites are just some examples where users create and publish content to be shared online. Among these there exist social tagging sites, where users are able to assign labels (tags hereafter) to resources other people publish or reference.

In this paper, we try to exploit the information web administrators provide on web pages (via web page meta-tags) as well as the wisdom of the crowds, that

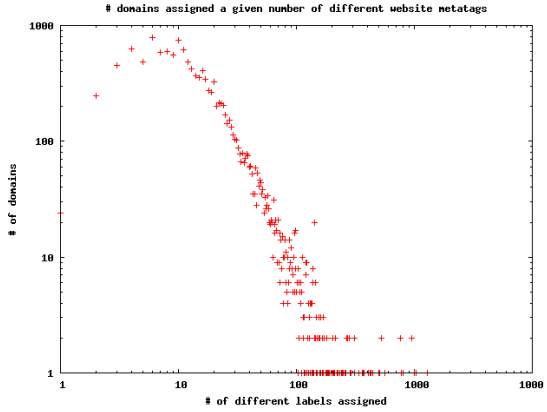

Fig. 3. Distribution of number of domains with N meta-tags assigned

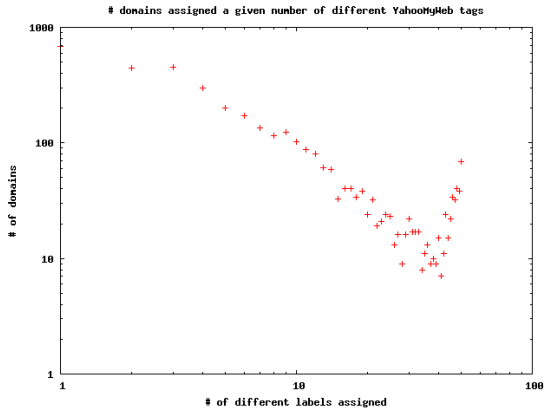

Fig. 4. Distribution of number of domains with N social tags assigned

is, the tags users assign to webpages on social tagging systems. In particular, we have built extractors that are able to receive meta-tags from any website and tags from systems such as YahooMyWeb. Table 3 describes the main characteristics of the data obtained through this process. It shows that the number of different social tags is still very reduced in comparison with the vocabulary used by web administrators. Additionally, only 37% of the domains visited by users were annotated by web administrators, and only 11% had any type of social annotation. It is interesting to see that social annotation already provides information from around a 5% of domains for which no meta-tags exist.

Figures 3 and 4 depict respectively the distribution of domains according to how many meta-tags they contain and the distribution of domains according to how many social tags they have been assigned in a logarithmic scale. Additionally, Fig. 5 shows the combined information also in a logarithmic scale, that is,

Table 3. Meta-tag and social-tag extraction statistics

Description	Amount	% over total
Different Tags assigned	108,851	
Different Meta-Tags assigned	102,188	
Different Social Tags assigned	12,497	
Total Tags assigned	321,533	
Total Meta-Tags assigned	281,090	
Total Social Tags assigned	40,353	
Total Domains with a tag assigned	14,436	41.49%
Total Domains with a Meta-Tag assigned	12,847	36.93%
Total Domains with a Social Tag assigned	3,839	11.03%
Total Users accessing a domain with tag	45,581	24.08%

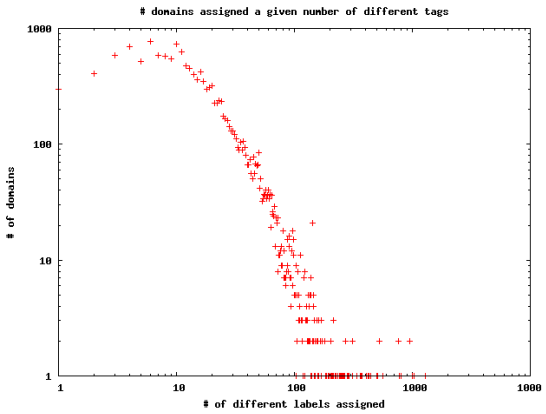

Fig. 5. Combined distribution of number of domains with N meta or social tags assigned

the distribution of domains for which we have any type of tag information. In Fig. 3 we can observe an inverse quadratic relation between the number of domains and the number of labels assigned, i.e. when the number of labels assigned to domains is small there is a quadratic increase, until 10 labels are reached, and after that the number of domains assigned a number of tags higher that 10 is quadratically reduced. Nevertheless, in Fig. 4 we can observe an almost linear reduction of social tags assigned, up to 60 tags, when there is a linear increase. That second part of the graph probably groups highly popular sites that are heavily tag by users. The combined information showed in Fig. 5 maintains the quadratic relation.

6 Combining Classification and Tagging Information

The combination of the two sources of information from sections above allows us to do two things: profile users both with categories and keywords for annotated pages and possibly predict categories of non-annotated ones based on

Table 4. Information available for domains and users visiting them

Description	Amount	% over total	Users	%
Domains with a category or social tag	15,949	45.84%	51,636	27.28%
Domains with category but no social tag	1,513	4.35%	14,350	7.58%
Domains with no category but social tag	10,466	30.08%	29,086	15.37%
Domains with neither category nor tag	3,970	11.41%	29,981	15.84%

the representative tags for each category. This section presents some results of the information we obtained from the combination of navigation logs, manual categorization of web pages and social tagging systems.

Table 4 shows some statistics about domains for which we can extract information based on the methods defined above and the number of users we are able to profile based on that (almost reaching a 30%, which already provides a very good amount taking into account the existence of many users accessing a very small number of domains or even seldom using mobile Internet). We also expect these numbers to increase when using information from other social tagging systems (e.g., delicious[5]) and more accurate navigation information (e.g., obtaining the full URL instead of only the domain). Additionally, we believe it would be good to perform the same analysis periodically in order to observe how the categories and social tagging systems adapt to the evolution of the navigation mobile users perform. Our intuition is that, since the trend for many users is to rely more and more on mobile Web navigation (especially after the emergence of flat rates in most developed countries), the coverage will increase, therefore providing more precise data.

Fig. 6. Tag cloud for category Science

Based on the information gathered, there is a subset of domains for which we have both category and tags assigned. We have used this information to explore the most representative tags assigned to each category. For this, we applied the Inverse Category Frequency (same as Inverse Document Frequency but checking the number of categories where a tag appears):

$$ICF_i = log \frac{|C|}{|\{c : t_i \in c|} \tag{1}$$

where ICF_i represents the Inverse Category Frequency of tag i, C is the number of top categories, and $|\{c : t_i \in c|$ is the number of categories for which the tag t_i has been assigned (through a URL). Basically, what this formula provides us is a classification of which tags are representative in order to characterize a

[5] `http://delicious.com/`

web page, based on whether the same tag is used for many different categories and are therefore not at all representatives[6], or whether it is used consistently only for one category[7].

Based on the results of this process, we were able to extract the most representative tags for each one of the top categories based on the known classified webpages. The interesting aspect of these observations is that we might be able to classify web pages for which no categorization data exists, simply based on the existence of "representative" tags.

To this aim, we selected all keywords with the maximum ICF, that is, they were assigned consistently to only one category and use all these tags to classify domains for which no category exists. The number of representative tags for each category are shown in Fig. 7, and Figs. 6 and 8 show some examples.

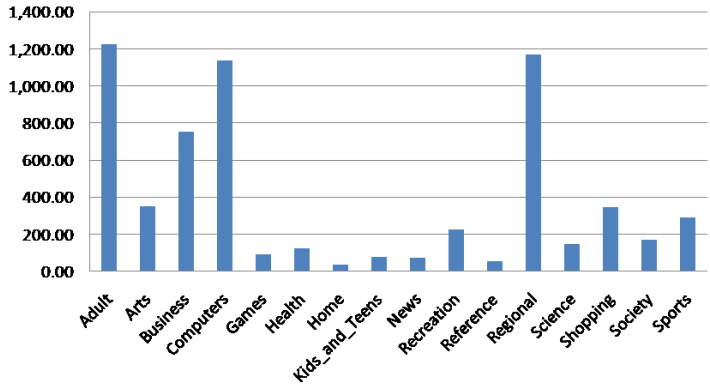

Fig. 7. Number of representative tags per category

Using these representative tags in our dataset allow us to classify a total of 5,617 new domains that were otherwise without category. This amount represents a 102,44% increase on domain classification with respect to the one made using only ODP categorization, therefore duplicating our coverage of domains, from 15,76% to 31,90%.

Fig. 8. Tag cloud for category Computers

With all the information collected, it is possible to improve and personalize the services provided to users. For instance, Figure 7 shows how adult content might be identified even if no previous knowledge of the site exists. This should

[6] Examples in our dataset include i.e. "noticias", "online", "free", "imported delicious", "internet", "photos" or "software".

[7] Examples in our dataset included "liverpool" or "pga" for category Sports, "blog" or "aol" for category Computers, or "science" for the category with the same name.

be combined with other approaches analyzing the content of the page in order to increase coverage, but even in that case, the tags extracted and classified as representative for adult content could be used as an input vocabulary for state of art approaches using text mining.

Additionally, user profiling is a key area for companies in order to enhance customer experience, and improve targeted advertisement and marketing among others. For instance, knowing the main categories a user is interested in (as our approach provides) allows to better select online recommendations on web portals or application stores.

7 Conclusions and Future Work

While very large amount of research has been dedicated to web profile analysis on the World Wide Web, and in the mobile world research has focused on navigation logs belonging to a web portal, there is to our knowledge no paper analyzing the navigation of users in a broader sense, having all their session activity independently of the portal or website they access. The advantage of this work is also that a handset typically corresponds to a single person, as opposed to desktop computers where it might not be the case.

This paper provides a first insight by analyzing the logs of three months of mobile web off-portal navigation (over 52 million accesses and 283,000 users) in a developed country and identified, among others, which type of web sites users accessed, which distribution these accesses followed or which main categories they are interested in. This information has also been matched with that of social tagging systems in order to better characterize users and categories, and the combination of these two approaches has been used to infer new categories for domains that were not previously classified.

The work shown in this paper gives an overview of the information contained in mobile navigation logs, but allows for much more advanced analysis. For instance, we plan to explore the usage performed according to time, location and demographic information of the users and see whether categories accessed vary on any of these dimensions.

Additionally, we hope to get navigation logs with the full URL (instead of only the domain) in order to better access categories and tags of visited pages, and also to include information of the web page content, as an additional input of data in order to try to improve the inference process. Since more data will be available, this will solve some already identified problems such as that of keywords appearing consistently in one category because of the use they are given, and not because of the real semantics. For instance, “Britney Spears” or “motel” were classified as adult, and “indoor” as sports because they were consistently used in those categories, but adding information from web page content might help increase accuracy. Also, analyzing co-occurrences of tags will help solve this problem, since there are tags with well-identified semantics (e.g., “pharmacy” or “drugstore” for health).

References

1. Albert, R., Jeong, H., Barabási, A.: The diameter of the world wide web. Nature 401, 130–131 (1996)
2. Huberman, B., Adamic, L.: Growth dynamics of the world wide web. Nature 401, 131 (1999)
3. Huberman, B., Pirolli, P., Pitkow, J., Lukose, R.: Strong regularities in world wide web surfing. Science 280(5360), 95–97 (1998)
4. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: KDD, pp. 538–543. ACM, New York (2002)
5. Jeh, G., Widom, J.: Scaling personalized web search. In: World Wide Web, pp. 271–279 (2003)
6. Shen, D., Chen, Z., Yang, Q., Zeng, H.J., Zhang, B., Lu, Y., Ma, W.Y.: Web-page classification through summarization. In: Sanderson, M., Järvelin, K., Allan, J., Bruza, P. (eds.) SIGIR, pp. 242–249. ACM, New York (2004)
7. Halvey, M., Keane, M., Smyth, B.: Mobile web surfing is the same as web surfing. Communications of the ACM 49(3) (2006)
8. Adya, A., Bahl, P., Qiu, L.: Characterizing alert and browse services for mobile clients. In: USENIX Tech. Conf., Citeseer, pp. 343–356 (2002)
9. Adya, A., Bahl, P., Qiu, L.: Analyzing the browse patterns of mobile clients. In: Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement, pp. 189–194. ACM, New York (2001)
10. Kamvar, M., Baluja, S.: A large scale study of wireless search behavior: Google mobile search. In: Proceedings of the SIGCHI conference on Human Factors in computing systems, p. 709. ACM, New York (2006)
11. Halvey, M., Keane, M., Smyth, B.: Predicting navigation patterns on the mobile-internet using time of the week. In: Special interest tracks and posters of the 14th international conference on World Wide Web, p. 959. ACM, New York (2005)
12. Anderson, C., Domingos, P., Weld, D.: Adaptive web navigation for wireless devices. In: International Joint Conference on Artificial Intelligence, vol. 17, pp. 879–884. Citeseer (2001)
13. Kamvar, M., Kellar, M., Patel, R., Xu, Y.: Computers and iPhones and Mobile Phones, oh my!, 801–809 (2009)
14. Timmins, P., McCormick, S., Agu, E., Wills, C.: Characteristics of mobile web content. Hot Topics in Web Systems and Technologies, 1–10 (2006)