Personalisation of Web Search: Exploring Search Query Parameters and User Information Privacy Implications – The Case of Google

Anisha T. J. Fernando
School of Information Technology and Mathematical Sciences.
University of South Australia
anisha.fernando@mymail.unisa.edu.au

Jia Tina Du
School of Information Technology and Mathematical Sciences.
University of South Australia
tina.du@unisa.edu.au

Helen Ashman
School of Information Technology and Mathematical Sciences.
University of South Australia
helen.ashman@unisa.edu.au

ABSTRACT

Personalised search adapts search results to the needs and interests of users. This is done through user data collected through various implicit and explicit methods and is used to build profiles of information needs of users. This paper highlights the need to explore search query parameters and determine their impact on personalisation. This is a first step in exploring the mechanisms of personal data collection and how personalised search uses personal data, which subsequently impacts the information privacy of users. It was found that location parameters have more impact on personalisation than the parameter 'pws' that switches personalisation on or off. Hence, it is important to undertake further research that investigates the impact of other types of search query parameters, their contribution towards search personalisation and their impact on user information privacy.

Categories and Subject Descriptors

H.3.5 [Information Storage and Retrieval]: Online In-formation Services—Web-based services

Keywords

Search Personalisation, Information Privacy, User Privacy Concerns, Search Query Parameters

1. Introduction

Personalisation of search aims to produce search results that individual users are interested in because it caters to their specific information needs. However, this raises privacy concerns given that personalisation collects personal user data to provide personalisation functions. Hence, it is important to investigate how personalisation occurs and what types of personal data is collected and used for the personalisation process.

The aim of the paper is to discuss a range of key privacy issues relating to web search personalisation. The scope of the paper is limited to discussing the initial progress of an experiment investigating location-based search query parameters. The paper is a first look at exploring the mechanisms of personal data collection and investigates the impact of selected search query parameter on personalized search results.

A brief background of search personalisation and the implications for personalised search is outlined. An initial experiment exploring the impact of search query parameters on personalisation is discussed. Future research directions with the aim of exploring the impact search query parameters have on personalisation and its impact on information privacy are also described.

2. Search Personalisation

Traditional information retrieval techniques are limited by the use of linguistic-based methods like keywords to express user information needs. Users have three broad categories of information needs according to their main goals, including navigational (to access a particular website), informational (to obtain relevant information on a specific area) and transactional (to execute a web-based activity) [1]. The information needs drive users to perform search queries and obtain relevant results. Users expect an instant and relevant response to their queries, whilst facing a highly dynamic web environment which is in a constant state of flux and the issue of information overload, where a countless amount of results are retrieved for a given search term [2]. These challenges led to the development of personalised search, which aims to provide relevant results to users based on their information needs [3].

Personalised search can be content-based or collaborative-based [4]. Content-based techniques use the content of items when determining relevant results that match user queries. Content-based techniques include:

- *contextual search* - where suggestions are provided based on the user's working context;
- *web search histories* - where search results are based on previous search history and selected results;
- *hypertextual data extracted from web pages* - where search results are modified by hypertextual algorithms to reflect criteria important to users;
- *rich user models* - where feedback is provided to actively build user models and store information about user preferences and search results;
- *adaptive result clustering* - where search results are grouped into clusters containing results on the same topic [3, 5].

Collaborative-based techniques use algorithms to produce results that are based on models of different users and their needs [6]. Collaborative-based techniques include *collaborative-based search engines*, which produce relevant results based on ratings of

prior users with similar preferences and *collaborative-community based recommendation systems* which produces results based on an analysis of community based search [3, 7, 8]. Both content and collaborative-based techniques collect information from users and build a user profile modelling user information needs and interests. This information can be collected explicitly through users providing 'relevance feedback' on search results or implicitly by capturing and processing click-through data [9].

Personalisation may be social-based using information from social media, location–based which focuses on geographic location details of users or may involve behavioural profiling and data aggregation based on longitudinal data collected [10]. This paper will focus on location-based personalisation.

Personalisation provides users with many benefits such as providing locally-relevant search results by catering to their user needs. In addition, it helps overcome the problems of information overload and ambiguity associated with using keywords to express user needs [11]. However, personalised search creates implications for privacy because personalisation requires the collection of user information to profile users and their information needs [12].

3. Information Privacy and Implications for Personalised Search

3.1 Information Privacy

Privacy is a social construct with different people having different attitudes towards privacy. Information privacy is concerned with personal information of users, which can subsequently be used to identify them [13]. In the web search context, users may involuntarily reveal information about themselves whilst searching and no longer have control over their data. Search engines collect large masses of this user data to profile user needs. User data may be public personal information, which is confidential, and non-intimate in nature or it could be non-public personal information [14].

It is pertinent to know the distinctions between non-public personal data and public personal data in the web search context. An example of non-public personal data would be a person's medical history that is held confidentially and stored securely and could only be accessed by a user with authorized access such as a medical practitioner. Public personal data would be a person's curriculum vitae containing work history on their personal website. In both instances there are general rules regarding access and this helps protect the data stored and uphold its accuracy and quality. Increasingly there are grey areas over what personal or non-personal information is and whether users (subjects about whom the data is stored) have enough access and control over what is stored or posted about them by others. Users may willingly divulge personal details online in a context where they feel comfortable to do so. However, most of what a user publishes online is seemingly stored forever given the countless backups and search result pages being indexed on a daily basis. Therefore, the 'right to be forgotten' in genuine instances is especially useful as it empowers users to have a stake or a claim if data posted about them is inaccurate or irrelevant.

In addressing public concern over privacy and personal information collected by organisations, governments all over the world are changing privacy laws to reflect the ever-changing and dynamic digital world we live in. In Australia, the Australian Privacy Principles embody 13 key ideals that organisations and government agencies that collect personal data have to abide by. These include open and transparent management of personal information, anonymity and pseudonymity where possible, guidelines regarding the collection of solicited personal information, dealing with unsolicited personal information, notification of the collection of personal information, use or disclosure of personal information, direct marketing guidelines, cross-border disclosure of personal information, adoption, use or disclosure of government related identifiers [15]. In addition, quality and security of personal information must be upheld and access to and correction of personal information should be provided. The European Union's Data Privacy Directive manifests user data protection principles similar to that of the Australian Privacy Principles. Personal data must be collected and processed in a legitimate manner, with explicit user consent obtained and where the individual can refrain from providing personal data for processing in applicable situations [16]. Based on the Australian Privacy Principles, in the web search context Australian users should be in control of their personal data. The user's personal details gathered through search queries may be used for purposes other than the original reason for data collection and this is not in accordance with national privacy laws.

3.2 Implications for Personalised Search

To personalise search results, personal user information is required which is routinely collected and profiled [17]. Significant levels of personalisation can be created through processing basic user information [18]. A search profile is a history of search queries where each query in the profile consists of the username, the time of the query, the query itself and when applicable, the link the user followed after the query was submitted and each query has context [19]. This enables a user profile to be built containing valuable context information. With improved capabilities of technologies to capture information, easily transfer and disseminate information and analyse information, personal data used in such a manner may have a serious impact on a person's privacy. Google, for example, is said to collect user data automatically which are recorded in search logs when users type in search queries. These include information such as the search query, IP address, browser information, date and time of request, cookies that identify the browser, hyperlinks clicked, operating system, language, processor type, screen resolution and colour depth, active plug-ins etc. [20, 21]. Cookies alone may not be effective in personalising as cookies are browser-dependent and more than one user may use a computer or a user may use multiple computers, thereby an inaccurate or incomplete user profile may be created [20]. Alternatives such as 'flash cookies' (local stored objects) behave differently to normal cookies and can be managed independently of the browser through the Adobe flash security setting [22]. However the privacy implications are that these are harder to manage, being much less well-known and lacking control settings in most browser preferences.

Personalisation may occur through account sign-in without account sign-in methods. For instance, from Google's perspective, an individual user's search history is used for personalisation when logged into a Google account. For users who are not using

or logged into a Google account, personalisation still occurs through cookies connected to a web browser and may remain there for a period of 180 days [21]. Therefore, it could be plausible that users who do not wish to have personalised results may still be given personalised results through use of such persistent cookies even if they proactively avoid personalisation through account sign-in by not logging into accounts. This increasingly difficult to avoid as users may use services from a suite of products that an organisation has, such as Google Search and Google+. It creates an exhaustive information source upon which to profile user actions and preferences. This indicates a key issue where users may be presented with search results based on what the personalisation algorithm determines suitable which is manipulated from the data collected through use of the search engine's services. This phenomenon named serendipity or the lack of it called the 'filter bubble' has significant implications where users depend or trust the search results being presented to them instead of actively looking at lower-ranked results that may provide them with a more representative view [12, 23]. This is particularly pertinent given the low levels of user awareness of what personalised results look like and how personalisation occurs and the use of personal user data in bringing about personalised search results.

This collection of user data by search engines also brings about a number of privacy problems such as: *aggregation,* where information collected about a person over a time can be combined to find out details of the person; *distortion,* where information collected in search query records may be misleading and may not reflect the actual intent of the users; *exclusion*, low levels of awareness by the public on what information is actually collected by search engines; *secondary use of data* which is not in line with the original purpose of data collection; and *political and social implications* of searching sensitive topics of interest [24]. One of the many benefits of search is that users can find out information as and when they require it and with personalised search, it aims to provide results that users are interested in. However, privacy is usually traded off against the capability to use functionality such as search [25]. This is because of the data collection that occurs and the ambiguity surrounding what exactly is collected. People should be able to actively control personal information and know how non-personal information about them is being used. Some key privacy requirements that should be upheld when data is transmitted include transparency, openness, notice and consent, where users are provided with options to control the level of personal data being transferred and the level of personalisation they prefer.

A key concern is that users may not be aware that their personal user data is collected and how it is being used [19]. Users may have also little control over how their personal data is being used and how they may retrieve and eliminate personal information [14]. Users may not have an actual choice in using search engines, as opting not to use search engines results in the inability to search and retrieve information [26]. Also the privacy policies of search engines are usually concise to avoid lengthy details, but in doing so may be vague and cover a broad spectrum of areas. However, it is ascertained that in such a context, people do not know what they are agreeing to and as such privacy policies fail in adequately communicating how user data is captured and used [26]. Also search engines are careful to avoid using terms like user profiling in their privacy policies [20]. Therefore, this may create a false sense of security in enticing users to opt in, or perhaps to not opt out.

Many different types of data in isolation are not particularly personal, but taken together can reveal many details about a person. In particular, when a collection of otherwise innocuous data is focused around a single user, a comprehensive profile of the user can be built up - this is linkability of personal data. A well-known example is that personal details about AOL user Thelma Arnold were ascertained from seemingly random search queries like searches for people with the last name 'Arnold' and 'landscapers in Lilburn, Georgia' she had made, after AOL released search keywords used by more than half a million users over a three-month duration. This was possible because AOL released query data pseudonymising each user with a unique numeric identifier assigned by the search engine, but neglected the linkability aspect by leaving the pseudonym unchanged for the entire three months of search data [25]. These search queries provide data that is used to build a user profile where decisions can be detected or inferred from the context. If these search queries are aggregated, then the search engine has the capability to identify and use personal details about people's lives.

It is also relatively easy to identify a person from information that is already public, but is supposedly de-identified. For example, William Weld, a former governor of Massachusetts was identified using only the ZIP code, birth date and gender from a combined pool of 'anonymised' medical data sold by insurance companies and publicly available voter registration data [27]. Therefore, by linking different sets of seemingly anonymised data, identities of people can be elicited and raises significant privacy concerns in ensuring anonymity of user information, especially when this data is publicly accessible.

The commoditisation of search to increase value-add for profit-oriented search engines resulted in targeted advertising. This behavioural advertising raises privacy concerns as it involves matching ads relevant to user needs based on the user profile, which captures user preferences and personal data from search history.

4. Exploring the Impact of Search Query Parameters on Personalisation and its Privacy Implications

In an attempt to further understand what types of personal data are transferred when searching, a series of experiments focusing on capturing and analysing HTTP requests and responses during these search requests are being conducted. An analysis of search query parameters is being undertaken because it is an effective method of examining the impact on personalisation and the ensuing impact on information privacy. It is important to investigate at this basic level because it is a valid measure of ascertaining what information is transferred when searching. These experiments will consider the different types of personalisation such as location-based, social-based, behavioural profiling and data aggregation and will investigate the impact across both widely-used search engines like Google, Yahoo, Bing and privacy-enhanced search engines like DuckDuckGo, Ixquick

and StartPage. As the investigations are in its early stages, the scope of this paper will discuss preliminary investigations pertaining to location-based personalisation and will be limited to using the search engine Google as an experiment vehicle. Google was chosen as the search engine as it is a popular choice for search and a majority of web search tasks are carried out using it. The overall aim of these evidence-based experiments is to explore the impact of search query parameters on personalisation and its subsequent impact on information privacy. Search queries were chosen from 5 different category types as queries are determined to have varying degrees of personalisation based on the category of search query [28]. The search query categories include technology, political news, entertainment, literature and science-related topics. These queries will be similar to everyday queries made by users and relate to topics that are popular and commonly searched about such as music concerts and political news. Personalisation has been shown to be more evident in specific categories of queries such as political and less evident in other categories like health related queries [28]. Through the use of an http-intercepting proxy tool, http and https requests and responses were captured when the search queries are performed and analysed to derive the impact of the search parameters.

When users search on the web, the HTTP protocol is used to send requests and receive responses [29]. These requests and responses consist of search query parameters like POST or GET parameters, HTTP headers or cookies. These parameters may capture or store user data that are transmitted when searching with the purpose of being relevant for personalisation and the potential for leaking personal data. Hence, HTTP headers, POST or GET parameters and cookies may act as potential sensitive data leakage points. Most search query parameters are not officially documented. Google's Privacy Policy highlights some examples of personally identifiable information, which may be used across the range of Google's services [30]. 'pws' is a GET HTTP parameter that can be switched on or off to control personalisation as widely described in the search engine optimisation community, but there is some concern over the impact of the parameter as a control over personalisation and little to evidence to prove it [31]. Hence, we assume that different parameters may have more significant impact on parameters than others.

An initial experiment was conducted to explore the impact specific search query parameters have on other parameters. Using a web debugging proxy tool to capture the search requests and responses sent to the search engine, the personalisation parameter 'pws' was manipulated across the 5 different search categories. This experiment was run across various scenarios involving all possible combinations of pws/location and sign in parameters. This included scenarios involving both types of personalisation (i.e. personalisation with and without account sign-in, switching the 'pws' parameter on and off and through location anonymity. This was then repeated with a gap of 15 minutes to minimise the carry-over effect (where conducting similar search queries to determine the effect a prior search has on the current search) across the 5 distinct search query categories [28]. To account for biases, each search query was done on a fresh instance of the web browser after having cookies and other persistent web data cleared after each web session. The test environment was constrained to be Windows 7 OS and Firefox Therefore, personalisation through account sign-in was through a Google account. In addition to the above scenario, an anonymising proxy, Tor[1], was used to provide location anonymity by spoofing the IP address.

Interestingly, there were negligible visible variances between the personalisation parameter 'pws' being explicitly switched on and off, or by default (i.e. 'pws' being absent from the search query) even with account sign-in or without account sign in (Figures 1, 2 and 3). Almost all of the search results were constrained to the Australian context. However, when Tor was used and the same search scenario was repeated, there were visible differences in the search results; for example, in one instance Google Germany was used with a mixture of results from UK, US and Germany to name a few (Figure 4). Hence, location parameters are observed to have more significant impact on personalisation than the personalisation parameter 'pws', which is designed to set personalisation on or off. This level of personalisation could vary depending on users with active logins and search history. Therefore, this initial exploration into whether there are differences with how parameters influence personalisation opens up an avenue for further exploration into the importance of search query parameters and identifying its influence on personalisation. On-going and future experiments will be refined to control for sources of noise such as *search index updates* – where search indices are updated on a regular basis, *distributed infrastructure* – where results may differ to data centres being located in different geographic areas, *geolocation* – where a user's IP address is used to produce results that are locally relevant, *a/b testing* – which is periodically conducted by search engine organisations to determine clickthrough preferences [28].

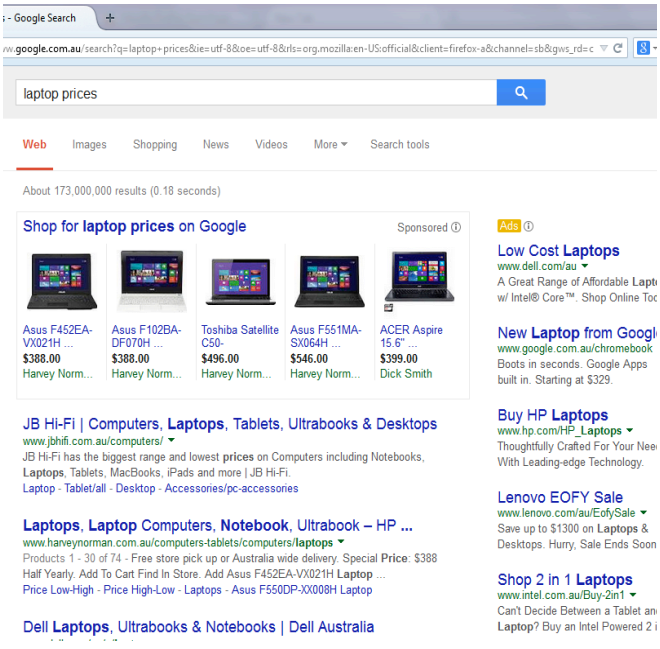

Figure 1- Example of a search result without account sign-in, personalisation parameter switched off and location spoofing off.

[1] Tor: https://www.torproject.org/

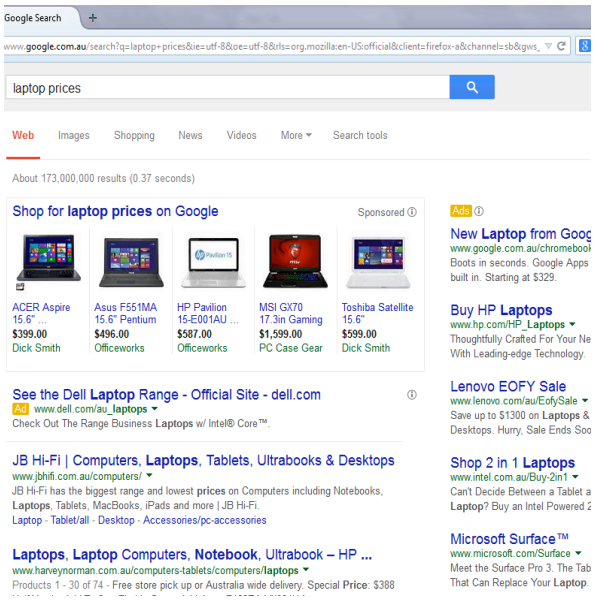

Figure 2- Example of a search result with account sign-in and the personalisation parameter on without location spoofing

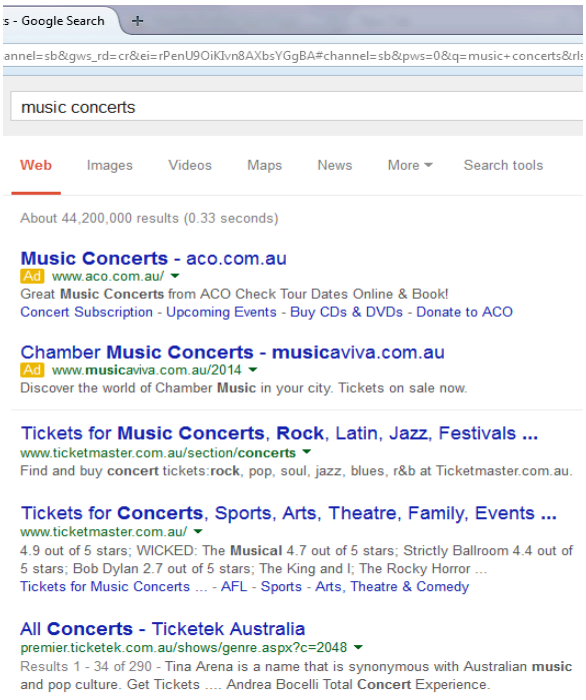

Figure 3 – Example of a search result with the account sign-in and the personalisation parameter and location spoofing switched off

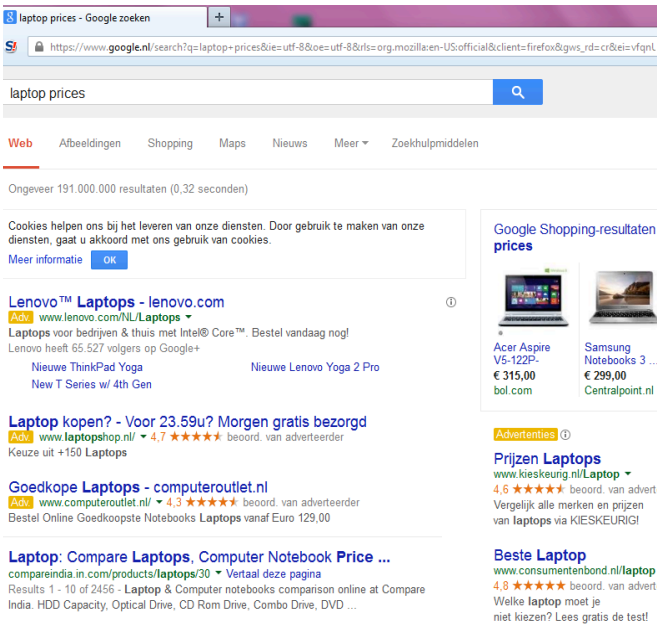

Figure 4- Example of a search result with location spoofing using Tor and the personalisation parameter switched on

5. Conclusions and Future Research Directions

It is important to find out the impact search query personalisation parameters have on privacy as well. Future experiments will focus on investigating the impact of other location-specific parameters, and its purpose (such as if it is being used for personalisation or for data collection), which would help examine the impact these parameters have on the information privacy of users by identifying what kinds of personal data is transferred. These future experiments will be automated by using software that can analyse and collect search engine results with the goal of investigating different parameter sets. The initial investigations provided familiarisation with how search query parameters work and established that certain parameters like location impact personalisation more than others. This impact could be due to a lack of extensive user search history available to the search engine in the experiment conducted. By identifying what types of user data are affected based on the identified search query parameters and its purposes, it would help recognise the significance of data submitted by users. Additionally, another key area to examine includes whether personalisation positively impacts the user experience and the reasons underpinning its impact. User privacy concerns could also be assessed in relation to personalised search and information privacy through user-based experimental studies.

With privacy being a major concern in the digital world, it is important to understand how personalisation works and what personal data are collected and used to perform personalisation. Identification of search query parameters and its purposes is a first step in determining how user information is used in the personalisation process. There is a need to conduct further research on the impact of the various types of search query parameters and determining its importance in influencing personalisation. Ascertaining what types of personal data is transferred by analysing search query parameters would allow a

clearer idea of the level of personal information disclosed and inform and validate the privacy concerns of users. Therefore, future research work will continue exploring the impact of various types of search query parameters, determine its purpose (if it is used for personalisation or data collection) and subsequently infer its impact on user information privacy.

6. References

[1] Broder, A. 2002. A Taxonomy of Web Search. *ACM SIGIR Forum.* 36, 2, 3-10.

[2] Lincoln, A . 2011. FYI: TMI: Toward a holistic social theory of information overload. *First Monday.* 16, 3, 1-15.

[3] Micarelli, A., Gasparetti, F., Sciarrone, F. and Gauch, S. 2007. Personalized Search on the World Wide Web. In P. Brusilovsky, A. Kobsa & W. Nejdl Ed. *The Adaptive Web*. Springer-Verlag, 195-230.

[4] Gao, M., Liu, K. and Wu, Z. 2010. Personalisation in Web Computing and Informatics: Theories, Techniques, Applications, and Future Research. *Inform. Syst. Front.* 12, 5, 607-629.

[5] Teevan, J., Dumais, S. and Horvitz, E. 2010. Potential for Personalisation. *ACM T Comput. Hum. Int.* 17, 1, 1-31.

[6] Shapira, B. and Zabar, B. 2011. Personalized Search: Integrating Collaboration and Social Networks. *J. Am. Soc. Inform. Sci.* 62, 1, 146-160.

[7] Reimer, K. and Brüggemann, F. 2006. Personalisation of eSearch Services – Concepts, Techniques, and Market Overview. In *Proceedings of* BLED *2006 -19th Bled eConference on eValues*, pp. 1-16.

[8] Smyth, B., Coyle., M & Briggs, P. 2011. Communities, Collaboration, and Recommender Systems in Personalized Web Search. In F. Ricci et al. Ed. Recommender Systems Handbook, Springer US, 579-614.

[9] Steichen, B., Ashman, H. and Wade, V. 2012. A Comparative Survey of Personalised Information Retrieval and Adaptive Hypermedia Techniques. *Inform. Process. Manag.* 48, 4, 698-724.

[10] Toch, E. Wang, Y, Cranor, L.F.: Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Model. User-Adapt. Inter*. 221,2, 203–220.

[11] Steichen, B., O'Connor, A. and Wade, V. 2011. Personalisation in the Wild: Providing Personalisation across Semantic, Social and Open-Web Resources. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, ACM, 73-82.

[12] Ashman, H., Brailsford, T., Cristea, A., Sheng, Qn.Z., Stewart, C., Toms, E.G., Wade, V. 2014. The Ethical and Social Implications of Personalisation Technologies for e-Learning. *Inform. & Manage.* 1-23. DOI=http://dx.doi.org/10.1016/j.im.2014.04.003

[13] Smith, H.J., Dinev, T. and Xu, H. 2011. Information Privacy Research: An Interdisciplinary Review. *MIS Quart*. 35, 4, 989-1015.

[14] Tavani, H. 2005. Search Engines, Personal Information and the Problem of Privacy in Public. *Int. Rev. Inform. Ethics.* 3, 1, 39-45.

[15] Office Of The Australian Information Commissioner. 2014. *Australian Privacy - Privacy Fact Sheet*. Australian Government, Canberra.

[16] European Union. 2011. *Protection of Personal Data*. European Union. Retrieved 26 April 2013. http://europa.eu/legislation_summaries/information_society/data_protection/l14012_en.htm.

[17] Berendt, B., & Teltzrow, M. (2005). Addressing users' privacy concerns for improving personalization quality: Towards an integration of user studies and algorithm evaluation. In *Intelligent Techniques for Web Personalization* (pp. 69-88). Springer Berlin Heidelberg.

[18] Krause, A., & Horvitz, E.2010. A utility-theoretic approach to privacy and personalization. *J. Artif. Intell. Resea.* 39, 633–662.

[19] Brandi, W. and Olivier, M.S. 2010. In Search of Search Privacy. In M. Soriano, S. Katsikas & J. Lopez Ed. *Trust, Privacy and Security in Digital Business.* 6264. Springer Berlin Heidelberg, 102-116.

[20] Aljifri, H. and Navarro, D. 2004. Search Engines and Privacy. *Comput. Secur.* 23,5, 379-388.

[21] Google. 2013. *Google Privacy Policy*. Retrieved 25 June 2013. http://www.google.com.au/policies/privacy/.

[22] Adobe. 2013. Analytic*s and on-site personalisation services - Adobe - Privacy Policy* Retrieved 7 September 2013. http://www.adobe.com/privacy/analytics.html.

[23] Pariser, E. 2011. *The filter bubble: What the internet is hiding from you*. Penguin Press, New York.

[24] Tene, O. 2008. What Google Knows: Privacy and Internet Search Engines. *Utah Law Review.* 2008, 4, 1433-1492.

[25] Zimmer, M. 2008. The Gaze of the Perfect Search Engine: Google as an Infrastructure of Dataveillance. In A. Spink & M. Zimmer Ed. *Web Search*, 14, Springer Berlin Heidelberg, 77-99.

[26] Nissenbaum, H. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford University Press. Stanford, California.

[27] Sweeney, L. 2002. K-Anonymity: A Model for Protecting Privacy. In International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 10,5,557-570.

[28] Hannak, A., Sapiezynski, P., Kakhki, A.M., Krishnamurthy, B., Lazer, D., Mislove, A. and Wilson, C. 2013. Measuring Personalisation of Web Search. In *Proceedings of the 22nd International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 527-538.

[29] Gourley, D., Totty, B., Sayer, M., Aggarwal, A., Reddy, S. 2002. *HTTP the definitive guide*. O'Reilly, CA.

[30] Google. 2014. Google- Privacy & Terms: Privacy Policy. Retrieved 10 June 2014. http://www.google.com.au/policies/privacy/

[31] Meyers, P.J. 2012. Face-off – 4 Ways to De-personalize Google. The Moz Blog. Retrieved 1 May 2014. http://moz.com/blog/face-off-4-ways-to-de-personalize-google