Taxonomy of Usage Issues for Consumer-centric Online Health Information Provision

Bang Viet Nguyen
Monash University

Frada Burstein
Monash University, frada.burstein@monash.edu

Julie Fisher
Monash University

Campbell Wilson
Monash University

www.manaraa.com

Taxonomy of Usage Issues for Consumer-centric Online Health Information Provision

Bang Viet Nguyen

Faculty of Information Technology, Monash University, Australia

Frada Burstein

Faculty of Information Technology, Monash University, Australia

Frada.burstein@monash.edu

Julie Fisher

Faculty of Information Technology, Monash University, Australia

Campbell Wilson

Faculty of Information Technology, Monash University, Australia

Abstract:

Consumers are increasingly using Internet portals when searching for relevant health information. Despite the broad range of health information portals (HIPs) available, usage problems with such portals are still widely recognized and reported. In this study, we analyzed usage data from an operational health information portal and identified ways in which these problems can be addressed. While previous usage data and log analysis research has focused more on user behaviors, query structures, and human-computer interaction issues, this study covers more comprehensive issues such as content. We describe a taxonomy of usage issues derived from a literature analysis. We describe how we validated and refined the taxonomy based on analyzing the usage data from an operational health portal. Findings from the usage data indicate that a range of content issues exist that lead to unsuccessful searches. The analysis also highlights that users' ineffective information seeking strategies are not well supported by the system's design. We use this taxonomy to propose a usage-driven, consumer-centered approach for dynamic improvements of HIPs. We also discuss the study's limitations and directions for future research.

Keywords: Health Information Portal, Usage Data, Information Retrieval, Log Analysis, Consumer-Centered Information Provision.

1 Introduction

Online health information provision is increasingly perceived as having a significant impact on consumers' healthcare, notably in providing effective healthcare knowledge, enhanced medical decision making, and, ultimately, improved public health outcomes (Cline & Haynes 2001; Keselman, Logan, Smith, Leroy, & Zeng-treitler 2008). Health information portals (HIPs) are considered a particularly effective and authoritative means for providing online content to health information consumers (Luo & Najdawi. 2004). (Hereafter, we use the term "health consumer" to refer to "health information consumers" for abbreviation purposes. In the context of using a system, we use the term "user" to describe the individual who directly uses the system).

HIPs, known as one-stop shops for quality online health information, are increasingly important for empowering health consumers (Williams, Nicholas, Huntington, & McLean 2002). In this research, we specifically consider user-centered HIPs (Burstein, Fisher, McKemmish, Manaszewicz, & Malhotra 2005). A user-centered HIP ideally delivers quality-controlled, user-sensitive content for health consumers (which we discuss in more detail later).

Previous literature reports that searching for health information in HIPs is more complex than simple domain-specific information searches (Zang et al., 2004; Keselman et al., 2008). Often highlighted is the mismatch between HIP functionality and the needs and skills of health consumers. It is widely recognized that health consumers have complex, changing, and heterogeneous information needs (Find/SVP, 1998). For example, health consumers may be searching for quite specific health advice or general advice, treatment or prevention information, information from experts, or community opinions.

Individual issues with seeking health information have been well studied. However, the larger picture of usage and how a better understanding of usage issues might lead to improved online health systems design have not been as widely researched. In this paper, we define usage issues as those that users directly experience when searching for health information. Therefore, in this study, we focus on examining usage issues from a broader perspective; in particular, those issues relating to the interaction between health consumers, health content, and online health systems. While acknowledging the commonality of online usage issues, we specifically focus our research on HIP usage issues given the increasing research interest around consumer-centered and usage-driven health information provision (Keselman, Browne, & Kaufman, 2008).

The move towards data-driven approach and consumerization of the Internet increasingly exploits the vast quantity of usage data to improve the service quality (Silvestri, 2010). We argue that online health information systems are no exception. We describes a usage data-based approach for both analyzing the issues and designing a HIP improvement strategy. We identify HIP usage issues through analyzing usage data that captured failed searches (understanding) and examine how to leverage these understandings to improve the design of health portals (use). Learning from usage promotes self-adapting and smart-learning capabilities for HIPs. For example, identifying the missing content searched for and other factors that lead to failed searches can help prevent situations that may lead to consumer dissatisfaction with a system.

This research involved three phases. First, we conceptualized a taxonomy of usage issues, which we drew from the e-health literature, that we considered most relevant to HIP context. Second, we conducted a usage data analysis, which provided further empirical insights on the issues and demonstrated how this data could be used to examine a range of usage issues. The second phase refined and confirmed the taxonomy. Based on the results, in the third phase, we were able to propose specific strategies for designing better, more sustainable, usage-driven improvements to HIPs.

More specifically, we address three questions:

1. What is the current thinking on key usage issues relating to HIPs?
2. What can we learn from usage data about usage issues in HIPs?
3. How can we use such knowledge to address the issues?

A preliminary version of this research was published as a conference paper (Nguyen & Burstein, 2013). In this revised paper, we provide a more comprehensive and detailed analysis of each individual issue with further illustrations from the usage cases. We also discuss the notion of usage gaps and usage-driven solutions in relation to the problem-solving strategy to meet health information-seeking requirements.

This paper proceeds as follows. In Section 2, we review the literature and highlight the major problems in providing consumer health information on the Internet. In Section 3, we describe the study's research methodology, which includes guidelines for analyzing and coding the usage data obtained from an operational HIP. In Section 4, we discuss the taxonomy and results of the usage data analysis, and, in Section 5, we proposed strategies for addressing identified usage problems in HIPs. Finally, in Section 6, we discuss the wider implications of this work; in particular, the opportunity to use the proposed strategies for improving and sustaining health information provision. Note that, while the case presented is related to the health domain, the outcomes may apply to other domains.

2   Online Health Information Provision

Online health information, especially in health portals, is often subject to tight quality management (Adams & Berg, 2004) and review processes, which pose many challenges in meeting health consumers' needs (Williams et al., 2002). Unlike a broad Web search, an individual who searches HIPs is more likely to receive a lower number of results because portals draw on a more limited pool of information resources. Failed searches often have a detrimental impact on health information seekers by leaving them feeling unsupported and confused (Kim, Lustria, Burke, & Kwon 2007; Evans, Manaszewicz, & Xie 2009).

Lack of understanding of usage issues can undermine the success of HIPs and has implications for both health consumers and information providers. Zeng and Tse (2006) note that health consumers may be confused about what remedial action can be taken in the event of a failed search, especially if the issues relate to the content or system, which can leave them with possibly ineffective corrective actions. For the information providers, ambiguity of usage issues creates challenges in development of effective, sustained solutions to address them (McCray & Tse, 2003).

Our research focuses on the concept of a user-centered HIP, which is an emerging model in the consumer e-health literature (Kwahk, Smith-Jackson, & Williges 2001). Despite being positioned in the field of consumer health informatics, it shares many characteristics with the patient-centered e-health notion that is described by Wilson (2009). Both concepts pay attention to HIPs' user-focused (by modelling content and services based on users' needs) and user-empowerment capacities. For example,as discussed by McKemmish, Manaszewicz, Burstein, & Fisher (2009) in terms of giving health consumers more control in personalizing and assessing the information quality).

Researchers have developed and adapted several theories in the context of consumer health searching. For instance Keselman et al. (2008) adapt the cognitive theory of information retrieval (Sutcliffe & Ennis, 1998) to explain how health consumers use online health systems to satisfy their information needs. We follow the approach proposed by McCray and Tse (2003) and focus on three major dimensions: content, users and systems. We note that the classical DeLone-McLean model of information systems success (DeLone & McLean, 1992, 2003) has also been extended to the e-health context in several studies (Van Der Meijden, Tange, Troost, & Hasman 2003; Yusof, Papazafeiropoulou, Paul, & Stergioulas, 2008) and that DeLone and McLean's work shares similar ideas on the determinants of an information system (IS) success (or conversely, failure), which assist in identifying factors behind the usage difficulties in HIPs. Notably, the three main factors of DeLone and McLean IS success model (i.e., information quality, service quality, and systems quality) are directly correlated with the notions of content, user, and system issues that we examine in this research. However, in our study, we adopted the constructs of McCray and Tse's model because their model provides more precise and specific interpretations in the context of search of health portals (for instance, service quality is arguably broader and may include factors beyond the way health consumers search).

Several researchers have analyzed usage data to improve health information searching. Herskovic, Tanaka, Hersh, and Bernstam (2007) and Scott-Wright, Crowell, Zeng, Bates, and Greenes (2006) examine user information needs based on the most popular topics searched by users of the PubMed and Medline Plus websites. Zeng, Kogan, and Ash 2002) and Zeng et al. (2004) analyze usage data and report several non-optimal searching issues, which echoes the results of the Eysenbach and Köhler (2002) study. However, Zeng et al. do not follow a systematic framework; hence, the issues they report arguably lack comprehensiveness. McCray and Tse (2003) describe a more comprehensive list of health searching issues, which we extend further in our research (Nguyen & Burstein, 2013; Nguyen, Burstein & Fisher, 2015). However, McCray and Tse only examine query data and, thus, provide limited evidence regarding content issues, which we have addressed in full as described in this paper.

As such, we can see that, while individual search issues have frequently been the subject of research in the area of health information, a more comprehensive study exploring the key usage issues in health search systems has yet to be undertaken. This gap is a concern when analyzing and evaluating HIPs, and may impact designing strategies to improve them. Furthermore, the literature focuses on either user issues (notably, issues on health information seeking behavior or search strategies) or information quality issues. To our knowledge, no other study has looked at all three: content, user, and system issues.

A research gap also exists in applying usage data to improve HIP design and content management. Prior health-domain research has employed usage data analysis mainly to explore issues with query terms (Zeng et al., 2004; Scott-Wright et al., 2006). To our knowledge, little work has been done to systematically explore the applicability (and the full extent of the limitation) of a usage data analysis method to study other types of usage issues; in particular, the content/user/systems spectrum in the context of health portals.

2.1   Usage Issues in Providing Online Health Information

In this paper, we focus on search issues related to the three major areas as McCray and Tse (2003) articulate: content, user, and systems. We summarize the literature on these categories below. In Section 4, we present our taxonomy of usage issues, which draws on the literature we discuss next and further elaborate on the impact of these issues in the context of online information provision for health consumers.

1. **Content issues:** Eysenbach, Powell, Kuss, and Sa (2002) identify several significant content issues for providing online health information that continue to be the subject of reported research (e.g., Pletneva, Cruchet, Simonet, Kajiwara, & Boyer, 2011). The key content issues frequently cited are the sufficiency and availability of content (Soualmia & Darmoni, 2005), permanence (O'Mahony, 1999), accessibility and readability (Berland et al., 2001), and permanence (or currency) (Eysenbach & Köhler, 2002).

    Sufficiency and availability is particularly important to online health information's sustainability (Soualmia & Darmoni, 2005). According to the HON Survey (see Pletneva et al., 2011), users rate availability of information as one of the most important factors, which confirms that content of health websites should be sufficient to meet their complex needs. Eysenbach and Köhler (2002) and Cline and Haynes (2001) report on permanence and highlight the importance of the "update cycle" of health content in ensuring content accuracy. Moreover, permanence is also important in the fluid, fast-changing environment of the Internet (Cline & Haynes, 2001). Information overload is often identified as a barrier in health searching, where too much irrelevant content may impact negatively on the user's experience (Burstein et al., 2005). Keselman et al. (2008) report on the broad spectrum of user health information needs and, hence, highlight concern with the diversity of online health content. Other important issues that have been identified with online health content include accessibility, readability. and the scientific complexity of content (Berland et al., 2001).

    In the context of health portal usage, the complexity of content issues is primarily attributed to the dependency on external resources and the controlled approach to managing the information content (Glenton, Paulsen, & Oxman, 2005)—particularly the involvement of domain experts (Evans et al., 2009). Evans et al. also maintain that, to ensure relevant, high-quality, and individually tailored access to health information, HIPs require domain expertise rather than automatic indexing to manage the resources. However, given that domain expertise is a limited resource, there are often sustainability issues in managing the task of identifying new content. This creates usage gaps, such as the differences between users' changing information needs and the content or between the indexing terms and users' search terms (Madle, Kostkova, Mani-Saada, & Roy, 2006).

    Quality is also an important issue related to content; however, this is a much contested and significant issue and was not in our scope.

2. **User issues:** these include the key topics relating to deficiencies in users' searching skills. Such key topics include language/medical query terms (Zeng et al., 2002), health literacy (Kogan, Zeng, Ash and Greenes 2001), sub-optimal searching strategies, and information processing skills (Zeng et al. 2004). The literature has also focused on the difficulty users have in expressing their heterogeneous and complex information needs, which systems do not often support (Josefsson 2006).

www.manaraa.com

Language/medical query issues describe the ineffective use of search keywords such as lay terminology, misspelt terms, or poorly formed terms (Zeng et al., 2002). In many cases, users do not recognize the issues and systems do not adequately support them (Zeng et al., 2002). Heterogeneous information is an issue if the website does not reflect the variety in information the user needs . For instance, Josefsson (2006) emphasizes the need to "manage various information needs related to different stages of the disease" (p. 10). Search strategy issues refer to "sub-optimal strategies" that users exhibit that can lead to bad outcomes, such as their ignoring useful search alternatives that retrieve better results. Recognizing and resolving ineffective search strategies, the web portals can retrieve more appropriate recommendations that lead to improved searching experience. Lastly, health literacy is an important issue for Internet health searching. Keselman et al. (2008) describes it as users' inadequate capacity to obtain, process, and understand health content. Such issues include difficulties in reading (e.g., when text is not readable) or vocabulary knowledge. Much has been written on the "information gaps" that hinder health consumers from understanding and consuming online health information (Alpay, Verhoef, Xie, Te'eni, & Zwetsloot-Schonk, 2009; Tang & Lansky, 2005).

3. **System issues:** we can group issues relating to systems into two broad categories: issues with the search/information retrieval functionality and issues with usability and the user interface (UI). Issues with retrieving health-specific information that Zeng et al. (2002) describe include the lack of user-friendly indexing terms or the lack of tools to deal with medical language (Zeng et al. 2002). Williams et al. (2002) highlight usability issues such as the difficulty of navigation and ineffective user interfaces for searching/browsing health information.

The user interface and usability issues that we include in this paper cover result overload, readability of results, medical jargon, and other usability issues such as navigation. Result overload in a health portal means that users must sift through large amounts of information, which is of particular concern when users have urgent information needs. Readability is important to health websites because, in previous surveys (HONSurvery, 2006), users have reported that easy-to-read text, appropriate text size, and information display are among the most helpful features when searching health information. Usage experience can be improved with clear presentation and explanations. Lastly, other usability issues such as navigation, site structures, complexity of tasks (e.g., number of steps required) are important to reduce the complexity of searching for health information (Marill, 2001).

IR-related issues include the deficiencies in health-specific information retrieval mechanisms and support for personalized/differentiated retrieval. Zeng et al. (2002) highlight some of the issues with online health information retrieval tools such as the lack of user-friendly term indexing and the integration of domain-specific knowledge (Zeng et al., 2002; Keselman et al., 2008). Tang et al. (2004) mention that health-specific search engines are not sophisticated enough to return more reliable and relevant results for users in comparison to general search engines such as Google. There is also a current limitation in providing personalized, differentiated information access (Josefsson, 2006), which allows users to effectively filter information for themselves and to improve the relevance of results.

3 Research Approach

We conducted the research in three phases. First, we conceptualized a taxonomy of usage issues, which we drew from the e-health literature, that we considered most relevant to HIP context. Second, we conducted a usage data analysis, which provided further empirical insights on the issues and demonstrated how this data could be used to examine a range of usage issues. The second phase refined and confirmed the taxonomy. Based on the results, in the third phase, we propose specific strategies and tools for designing better, more sustainable, usage-driven improvements to HIPs

3.1 Phase 1: Taxonomy of Usage Issues: Conceptual Design

In the literature analysis, we identified common problems and usage issues with providing consumer health information online. We used several comprehensive survey papers in the field of consumer health information seeking, such as Cline and Haynes (2001), Stavri (2001), and Keselman et al. (2008), as our starting point for identifying usage issues. We followed on with additional search from the journals and conferences in the fields of information systems, e-health, and information retrieval (in particular, health information retrieval). The general structure and key themes forming the taxonomy emerged from the

analysis. We created or group together major categories until no more new categories appeared. We reviewed and refined the taxonomy iteratively through this analysis.

We grouped the categories to provide a high-level overview of the issues, and we based it on several categories described in other papers (e.g., Cline & Haynes, 2001; McCray & Tse, 2003). We describe the results of this categorization in Section 4, Phase 1.

For instance, "query and language" only capture problems due to "faulty queries" and imprecise search keywords, a topic that Zeng et al. (2004) cover extensively. For another example, "low flexibility to relax search" relates to users' search effort (users could obtain more favorable search outcomes if they persist in their search sessions and varies their search strategy more broadly). In that case, we think it makes sense to include it in "user effort and persistency".

In surveying the literature, we needed to define the boundaries when reviewing papers on user search issues since they represent a large area of research in the field of user information seeking and information retrieval. We limited user issues to those that were prominent in the e-health field, such as those identified by Cline and Haynes (2001). These include the sub-areas of query and languages (medical search terms), searching behaviors on health websites, search strategies, and users' information-seeking skills. For content issues, we included several papers that centered on the quality of health information and the issues in content management. We also identified systems issues from these papers. Note that the grouping of usage issues into categories, while guided by the literature, is subjective.

We excludes several issues we considered to be too general to the area of HIP usage issues[1] from the scope of the taxonomy:

- Digital divide issues (high-level issues; e.g., inadequate access for the elderly or non-English speakers; Keselman et al., 2008)
- Accessibility (e.g., compliance with Web guidelines; Eysenbach et al., 2002)
- Interactivity issues (Cline et al., 2001)
- Relevance, precision, ranking of results (Kogan, Zeng, Ash, & Greenes, 2001)

Quality issues (given the size of the topic; Eysenbach et al., 2002)

3.2   Phase 2: Usage Data Analysis

In this phase, we:

- Assessed the problems empirically, explored to what extent problems we could identify problems by analyzing usage data, and analyzed their distribution.
- Validated and refined the taxonomy based on analyzing data.
- Explored the link between observed problems and possible problem-solving strategies to improve HIP usage.

We collected the usage data for this study from a specialized healthcare information portal, Breast Cancer Knowledge Online (BCKOnline) (http://www.bckonline.monash.edu.au), an Australian user-centered personalized health information portal for breast cancer (McKemmish et al, 2009). BCKOnline is a good model of a consumer health website because it provides quality-controlled health information to a diverse range of users including early/recurrent/advanced cancer patients, families, and caregivers, with each resource having been evaluated by a domain knowledge expert. The user-centered modelling process assigns each resource to different searchable categories (such as "scientific", "plain, detailed", or "personal stories"), which allows the user to tailor their search to their needs. BCKOnline provides three searching modes: simple search for a quick start (which is the most used mode), personalized search for tailoring information to users' specific needs by simply selecting some or all of the categories describing their situation, and topic search for browsing pre-formulated breast cancer topics. The simple search contains a query form with minimal searching options, while personalized search provides extensive searching options. Given the tailored-information approach, a search is considered problematic if it

---

[1] As we mention in Section 1, the notion of usage issues in this paper refer to issues experienced by users when searching in HIPs that are detectable when looking at usage data.

retrieves all of the results or no results, which we described in the coding rules below. Figure 1 shows a screen shot of the personalized search feature.

The original data source included over 300,000 entries. We captured 3582 search cases involving searching activities through the "personalized search" and "simple search" (we excluded non-search visits) between the period of 5 March 2008 to 31 August 2009. Note that "personalized search" is the most representative and comprehensive search option on the website. The difference between the two search modes is only the additional search options. Also note that we focused on identifying what the issues might be (and how we might detect them from log data) rather than to study the specific issues of each search mode. As long as a search failed, it was a subject of interest in the usage data analysis, regardless of search mode.

Of the 3582 searched cases, 300 were identified as failed ones (we define what can be treated as failed search below). We used these 300 failed searches as a sample for manual classification. Captured data included queries, user profiles (preferences), search options, search modes, click-through data (e.g., access to the result pages), search refinements, and search outcomes. Appendix A provides examples of captured data, and Appendix B presents coding process.

We also replicated search cases with relaxed search options to determine all relevant content for a given search. Since we replicated the search manually (with certain information being replicated automatically), we could reconstruct a relatively authentic view of the user search process. By combining this with associated information from the resource database, such as the amount of available content, indexing terms, and content distribution, we could replicate the issues holistically. Accordingly, the usage data analysis methodology validated a broader part of the taxonomy and provided additional focus to content issues and not just user-system interactions.
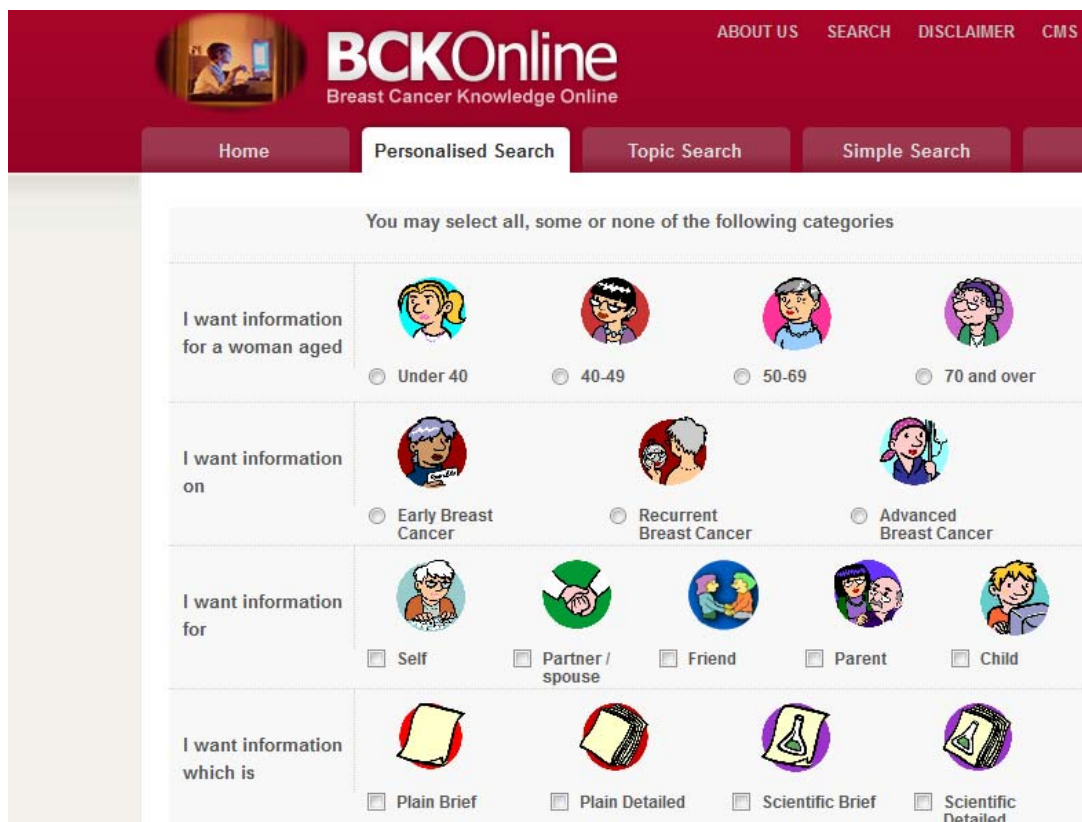
Figure 1. Screenshot of the Personalized Search Feature of the BCKOnline Portal

Also note that, while we coded log data, we replicated/had to replicate many data elements (i.e., did not exist as-is in the log). For instance, the search outcomes (number of search results) were not available in the original Apache Web log, so we had to obtain such information by replicating them (through a programmatic tool). Only with such information were we able to classify failed/non-failed searches). Once integrated, we referred to those data altogether as "usage data".

Analyzing usage data as a method is consistent with other research (Herskovic, Tanaka, Hersh, & Bernstam, 2007; Bernstam, Herskovic, & Hersh, 2009). We acknowledge that the method does have certain limitations, particularly reliability concerns, since users' needs and experiences can only be understood indirectly (McCray & Tse, 2003). However, the method has certain advantages that justify our choosing it. In particular, a usage-data based analysis allows us to approach the problems in a more comprehensive and non-intrusive way compared with other data-collection methods.

We coded the usage data based on previous research, such as McCray and Tse (2003). However, while much previous research has focused exclusively on analyzing query failures, we covered more types of usage data. Therefore, we were able to study a wider range of usage problems. When examining an issue, we were able to inspect both the properties of the content and users' searching behavior or the response of the system to determine the reason for the issue (no result or too many results). The usage-based approach is also in line with our chosen strategy for addressing problems in online health information (see Section 2).

To address reliability concerns, we employed several techniques and strategies to assist in the manual coding process.

3.2.1   Coding Procedure

To identify usage issues and problems, we sampled problematic use cases, which we classified as follows: (1) no/few results from the search, (2) too many results (broad searches that returned all results), and (3) abandoned searches (the user left after a single search). Appendix B provides a diagrammatic summary of the coding process. We used the following operational definitions: "no/few results" (less than five results; i.e., less than a page of results) and "too many results" (where more than 400 of the total content (800+ resources) was returned.

We deduced the nature of search failures based on observing the user's intent as expressed by the query, users' interaction with the system after the results were retrieved, the amount of effort from users, and the search outcomes. Table 1 summarizes the strategies we used to identify and code the issues based on the usage data analysis.

Table 1. Strategies for Coding Usage Data

Usage issue	Usage data analysis technique
User issue	Question asked to identify this issue: "Could the search be improved if the user had tried alternative ways of searching?". We examined the user's effort (number of search actions), strategy (whether the user employed different search modes or search options), and search queries. We tested each user's search strategy as follows: (1) alternative query terms, (2) different search modes, (3) different search options. Example 1: a user searched for the keyword "pill" (a very general and "layman's" term). followed by the more specific phrase "hormone pills after breast surgery". Therefore, we coded the first search as a "layman's term" (U1) issue and a broad query (U7) issue. While the user was able to rectify the search in this case, the issue was still recorded (with the first search) to capture the phenomena behind the first search (which another user might encounter in the future). Example 2: a user searched for "chemotherapy" (a common topic with available content). However, the user obtained no results because of the very narrow (i.e., specific) search options ("under 40", "early breast cancer", "medical type", "plain-detailed format"). The user did not make any attempt to relax the search (change to a simpler search option), which could have resulted in retrieving some content. Therefore, we coded this search as a "user strategy" issue (U4). Note that we also recorded a C6 (lack of content diversity for this specific user demographic) because this is a legitimate information need. It highlights that there could be multiple issues leading to a failed search (the system does not have content catering for this user group, and the user might not have been flexible to change the search). What is more important to highlight is that these are all potential issues, so systems need to be designed to respond to both (e.g., to prompt users to change to a simpler search option or to add content for this specific needs if it warrants it).

Table 1. Strategies for Coding Usage Data

Indexing issues	We defined these issues broadly as: the content existed (found by relaxing the search), but the search failed because of keywords or indexing options Example: there was content indexed with "tamoxifen" but a user searched for "taxol". Taxol is acceptable as a term in the MESH glossary (http://www.nlm.nih.gov/mesh/MBrowser.html). It is a potential indexing issue because the alternate (but legitimate) term has not been added to the index. Example: no results were found for the query "brachial plexus metastasis" with "medical type" search option, even though "medical-type" content existed for this topic. The issue here was with the metadata indexing ("medical" type) that the domain experts might have assigned incorrectly (as "supportive" instead). This was flagged as an issue to highlight the potential problems with the indexing process, which, once detected, can lead to a review mechanism to avoid such problems from occurring.
Content (terminology) issues	The query in MESH (a medical controlled vocabulary at The National Library of Medicine) was checked to determine whether a search contained medical or scientific terms. Medical language issues could then be checked against the portal's glossary database. Example: a user searched for "chording", which retrieved no content. A check in MESH found that this was a query issue because the correct query should be "cording".
Content issues (missing content)	The search was relaxed to see if there was any relevant content. Initial search options were removed, and more specific keywords were removed to broaden the search. Example: the initial search "pregnancy and breast cancer" produced no results. We changed the search to "pregnancy" to see whether that would retrieve any content and found that content for the "pregnancy and breast cancer" topic was missing.
Classification/ indexing issue	To decide whether a query was in scope, we searched Google search and examined the results to see whether they were relevant to the context (breast cancer or cancer). Example: a user searched for "secondary lung", which was probably out of scope because it could not be found in the MESH glossary.

We developed a multi-faceted data analyzer tool to integrate different aspects of usage data to facilitate the coding. We coded and classified each usage case manually based on the nature of the issue. We also noted qualitative comments on some cases. We could observe and assign multiple issues to some usage cases (see Section 4.4 for details on the tool and the examples).

As much as possible, we used objective empirical evidence to analyze the data. However, because we focused on discovering the phenomenon behind an issue and on aligning it to the proposed taxonomy, we focused on the technical precision of the coding. We also checked whether any other issues existed that did not fit the proposed taxonomy and extended or refined the taxonomy accordingly.

3.3   Phase 3: Applying the Usage Issues Taxonomy to Improve HIPs

We developed a framework of usage-driven improvements for HIPs. We define usage-driven improvements as those that are based on insights obtained from historical usage data. Drawing on the empirical findings for how each issue could be addressed and the insight that usage data provided to address the issue, we propose a set of solutions to assist in managing HIPs' content to identify the issues and to prevent these issues from occurring.

We describe a usage reporting tool to illustrate how insights derived from the usage data can be used to address the problems.

4   Results

In this section, we present the results of our three research phases as applied to the BCKOnline portal data.

4.1   Phase 1: Taxonomy of Usage Issues: Conceptual Design

We analyzed the literature to identify the key usage issues faced by users of health portals. We coded and classified each identified issue under one of three areas: issues relating to content, issues relating to how users searched, and issues relating to the system. Table 2 presents the outcomes of the literature analysis and the supporting sources. Note that we do not list all papers drawn on at this stage of the research. Table 2 contains the key references and those representative of the literature.

Table 2. Taxonomy of Usage Issues

Issues	Categories & Description	Supporting Research
Content issues (C)	C1. Insufficient content for a specific health information need	Cline & Haynes (2001) Kogan et al. (2001)
	C2. No information category or a general content topic	Soualmia & Darmoni (2005)
	C3. No permanence or maintenance of content	Cline & Haynes (2001)
	C4. Overloading of irrelevant or low-quality content	Burstein et al. (2005)
	C5. Misclassification/incorrectly indexing of information content (e.g., misuse of terminology in indexing)	Kogan et al. (2001)
	C6. Lack of information diversity to accommodate the heterogeneity of health information needs	Josefsson (2006)
	C7. Scientific complexity in content	HONSurvey (2006)
User issues (U)	**Query & language issues**	
	U1. Layman's terms or inaccurate scientific query	Zeng et al. (2002)
	U2. Misspelling	McCray & Tse (2003)
	U3. Out-of-scope query	Zeng et al. (2004)
	Search Strategy	
	U4. Single search strategy	Benigeri & Pluye (2003)
	U5. Confusion with use of search tools or query composition	Keselman et al. (2008)
	U6. Use of overly-scientific or medical terms	Zeng et al. (2004)
	U7. Use of broad terms that are too general	Zeng et al. (2004)
	U8. Limiting search with narrow options or narrow topics	Zeng et al. (2004)
	User effort and persistency	
	U9. Low persistence in searching	Zeng et al. (2004)
	U10. Low flexibility in relaxing searches	Zeng et al. (2002)
	User Information Skill	
	U11. Mental model, unclear intent when expressing needs	Keselman et al. (2008)
	U12. Consumer health literacy	Kogan et al. (2001)
	U13. Misunderstanding results, search abandoned	Keselman et al. (2008)
System issues (S)	**User interface**	
	S1. Result overload S2. Readability of results	Fisher et al. (2009) HONSurvey (2006)
	S3. Medical jargon used without language tools	Fisher et al. (2009)
	S4. Usability of health website	Marill (2001)
	Search functionality	
	S5. Limited search functionality	Tang et al. (2004)
	S6. Complexity of search tools without proper explanation	Williams et al. (2002)

In the following sections, we elaborate on these issues in the context of providing consumer health information.

Drawing on the issues identified in Table 2, Figure 2 depicts the main themes of the conceptual taxonomy of usage issues. The broad categories are based on those issues commonly mentioned in the literature. For instance, we group several user issues that describe problems with medical query languages under "query and language".
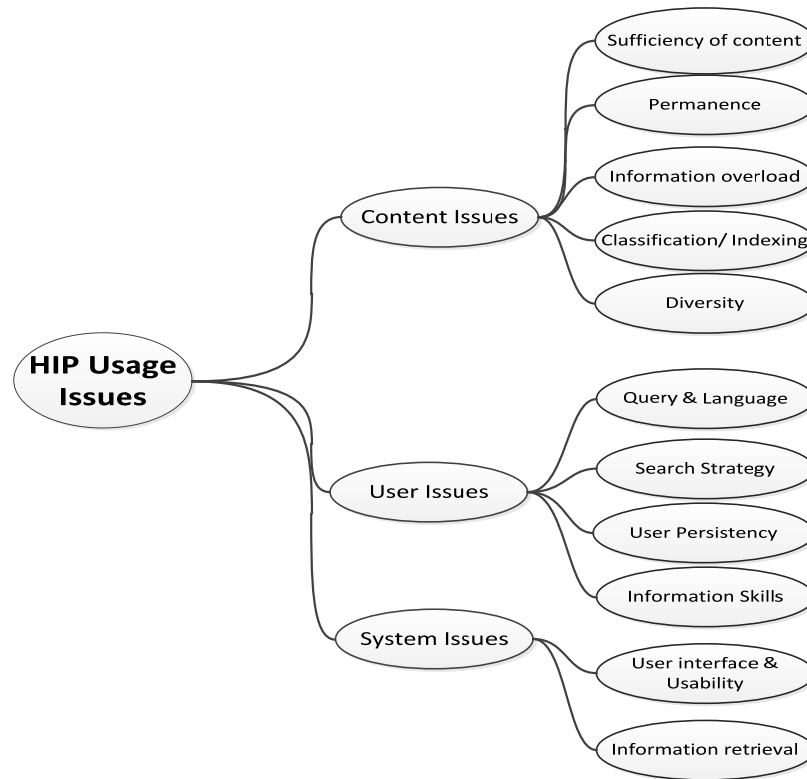

Figure 2. Key Usage Issues (Nguyen & Burstein, 2013)

4.2 Phase 2: Usage Data Analysis

From analyzing the BCKOnline usage data, we identified more specific usage issues that we subsequently used to further refine the taxonomy. First, we report the findings for each individual issue. Based on the proposed theoretical perspective, we also highlight two major gaps at the intersection between these usage dimensions. These include user/content and user/systems gaps. Subsequently, we discuss the examples for each of the three usage issue categories presented in Table 2. We present the results as a percentage of problems identified through the usage analysis for each of the codes defined in Table 2. For each issue, we also suggest a usage-driven strategy to improve it in a HIP.

4.2.1 Content Issues

We can classify the content issues into six categories (see Table 3). The table also contains the relative distribution (in percentages) of these issues as presented in BCKOnline data. Note that a usage case might be classified with more than one issue; hence, the numbers do not add up to 100 per cent.

Table 3. Percentage of Content Problems

Issue	Percentage (n = 300)
C1: Lack of specific content	25.30%
C2: Lack of general content	19.00%
C4: Information overload	10.70%
C5: Indexing issues	37.30%
C6: Information diversity/ balance issues	26.00%

4.2.2 Lack of Content

Issue: while the literature reports a lack of relevant online health content in general, we suggest that it may be beneficial to distinguish between the lack of content for a specific information need (C1: 25.3%) and that for an information category (C2: 19%).

Examples: searches for treatment cost/finance-related information emerged from our analysis without any content to match, and no existing category existed in the portal content for this type of information. This may prompt the addition of a new category to address these needs as opposed to adding specific missing documents to an existing content category (in the case of C1).

Improvement strategy: finding effective automatic techniques to differentiate between C1 and C2 will benefit organizations in developing and maintaining their HIP's content to better meet user needs.

4.2.3 Information overload

Issue: the overload of irrelevant content (10.7%) was apparent with several searches, especially searches with broad queries or that ineffectively filtered information via search options. We found evidence that this issue was associated with other issues, such as query formulation strategies and the users/systems gap in understanding the retrieval functionality (e.g., a false assumption of the robustness of the search). We also identified cases of excessive content (e.g., too much content for a single result).

Examples: we often saw broad queries without a precise context, such as "treatment", "cancer", and "risks". Users also seldom subsequently revised their searches, and the system didn't provide them adequate support (e.g., disambiguation).

Improvement strategy: mining these failed searches from usage data can improve query recommendation/ refinement tools. Users should also be presented with alternative recommendations, such as those with a more precise context.

4.2.4 Classification/Indexing Issues

Issue/problem: among the most prevalent content issues were those relating to indexing (37.3%), which is often sophisticated in health searches. Several cases in our data showed that, while the content existed in the database, the indexing may have prevented users from finding it. We observed several sub-issues, including the level of specificity, inclusiveness, and relevance of indexing terms. We also noted categorization issues (e.g., how a document resource is assigned to a subject heading) as potential issues and the assignment of meta-data to documents (e.g., which disease stage is suitable for a given document).

Examples: user-oriented search terms such as "hook wire biopsy" were not accounted for in the index, although the content topic (biopsy) existed in the system. Searches such as "self-examination" (relating to diagnosis) could have return indexed content with different meanings (e.g., "examination of the report showed that…")

Improvement strategy: the findings demonstrate the necessity of having usage-driven indexing systems that are consistent with users' search terms. This requires more inclusive and more accurate indexing systems and an ontology that covers user-generated terms.

4.2.5 Diversity/Imbalance Issues

Issue: the diversity of information (26%) is an issue for providing user-centered health information as reported in the literature. For instance Josefsson (2006, p. 10) argues for the need to "manage various information needs related to different stages of the disease". Our findings confirm this issue. For instance, the content distribution may not be balanced across different categories, may not be representative, or there may be a lack of content for a particular audience's profile.

Examples: no content for the search "breast reconstruction" in the medical-only category existed. The problem was not obvious to the domain experts who oversaw the database, although content relating to this topic did exist in another category (hence, this is not a "lack of content" issue per se).

Improvement strategy: usage issue reporting needs to highlight the diversity/imbalance of content problem effectively. Domain experts responsible for maintaining HIP content need information on the

content topics that are not well represented. Topics that are under-represented (with respect to real usage) might be subject to further reviews.

4.2.6 Other Issues

We did not replicate the issues of permanence and expiry of content (C3) and scientific complexity of content (C7) due to the limited data in our data set. Cline and Haynes (2001) raise lack of permanence as an issue of health information and refer not just to the expiry of the content but also to the changing or moved status of the content. Analysis of usage data allows one to detect those problems through, for example, examining users' "bounce-back" behavior after clicking a URL.

The level of scientific complexity in health content is an issue that is associated with health literacy, which researchers have described as one of the major difficulties for health information consumers (Keselman et al., 2008). With the usage analysis, we could replicate the searches; however, we did not examine the content's technicality. We argue this issue can potentially be addressed by conceptualizing the notion of "scientific complexity" in health content and including the prevalence of medically inclined keywords or technical jargon, among other factors (Keselman et al., 2008).

4.2.7 User Issues

User issues highlight the deficiencies in the search strategy, the choice of query, and the level of persistence and users' information processing skills. The findings also suggest that we need a range of support tools to address those searches that do not perform well and result in user issues. Table 4 presents the distribution of issues present in BCKOnline data.

Table 4. Percentage of User Problems

Issue	Percentage (n = 300)
U1: Use of layman terms	3.7%
U2: Misspelling	3.3%
U3: Out-of-scope search	5.3%
U4: Single search strategy used	28.7%
U6: Overly scientific query	18.7%
U7: Broad or general terms used	29%
U8: Narrow search options	30.3%
U9: Low persistence	N/A (reported via averaged session length)
U10: Low flexibility	14.7%

4.2.8 Query and Language Issues

Issue: through log analysis, we found evidence for all three issues related to language as identified in through the literature analysis (i.e., usage of layman terms, misspelling, and out-of-scope queries). Although they are not prevalent, they still constitute a noticeable percentage of the failed searches. Replicating the searches demonstrated the shortcoming of the current system in supporting these issues.

Examples: use of "chemo" (a layman's term) instead of "chemotherapy", "angela", "booster tower" (unclear on the scope of these queries; they are potentially out of the breast cancer domain. The users ended the search without further refining them or without the system offering suggestions.).

Improvement strategy: identify common query faults/out-of-scope queries in usage logs to better train and improve language support tools (such as "did you mean?" function common in online dictionaries/thesauri) and to augment the content's scope.

4.2.9 Search Strategy Issues

Issue: the usage analysis clearly evidences these issues. A substantial number of cases involved short and simple strategies (U4: 28.7%) that were not effective. These included the search mode or choice of broad/narrow query terms that do not retrieve any results, which suggests that users might not be aware

of alternative options to change the search effectively. The usage analysis suggests that, in health searches, the use of overly scientific keywords (U6: 18.7%) or broad/general terms (U7 - 29%) are common factors contributing to failed searches.

Examples: a term such as "grade invasive ductal carcinoma in situ" did not retrieve results. Another user used "surgery" followed by "breast construction" (both are general terms that did not help refine the search). Other observed examples include the continued use of a particular search mode or search options that failed to retrieve results.

Improvement strategy: recommendations pertaining to search strategies may be needed to broaden users' search space, especially when it comes to dealing with a complicated health information need and user's poor/nonexistent awareness of alternative search options. These include suggestions for alternative search modes, search tips, or different ways to formulate queries (such as broader or narrower terms).

4.2.10   User efforts and persistency

Issue: the literature reports a lack of persistency in searches, especially when searches failed (Scott-Wright et al. 2006). Our usage analysis echoes this finding, with an average of 2.1 queries per search session. In many cases, users exhibited inflexibility in changing and so did not pursue the search.

Examples: many cases of abandoned searches after a single query.

Improvement strategy: using situation-aware supporting tools that encourage users to extend their searches may eventually achieve better results, which is particularly important in certain scenarios, such as repeated failed search trials or abandoned searches, where users may feel frustrated. These scenarios can be identified from usage logs.

4.2.11   Issues Relating to Users' Information Skills

Issue: the issue of unclear intent (referred to by Zeng et al. (2004) as the mental model) is partially observable through the queries data, while the other issues involve the perception of users these cannot be identified in usage data.

Examples: health information needs are complex with many aspects that need to be described. For instance, a search for "massage" could refer to many aspects of massage (e.g. the methods or the benefit of massage).

Improvement strategy: the system should support users in expressing and disambiguating their information needs. Such a feature could be developed from usage data.

4.2.12   System Issues

Addressing system issues is crucial to the success of online health systems, particularly those pertaining to usability, system design, and information retrieval. Fisher, Burstein, Manaszewicz, and Lazarenko (2009) assess system problems, including features such as personalization, user interface, and search functionality, and suggest improvements for the user search experience. We do not discuss systems issues because they fall outside our scope, with the exception of issues relating to content management that might lead to content issues (such as lack of content). We include system issues in the revised taxonomy because of their importance and for completeness; however, the usage data analysis does not clearly evidence them. Nevertheless, note that we observed several shortcomings of the system when replicating the usage cases. Some of the implications and improvement strategies suggested above may also address system issues.

Co-occurrence of Issues: Who is Responsible?

Often, we could single out an issue category based on the usage data. Problems may occur due to factors with the content (or the system) and the users. Search failure is amplified when users experience deficiencies in the content or the system. At the same time, a lack of skills or strategies to identify and rectify the problems can occur. In this study, we noticed that there were some notable co-occurring issues as observed in BCKOnline data. The co-occurrence issues sufficiently indicate that many failed searches often involve both content and user issues. However, note that, because of the limited scale of the qualitative data coding, we cannot generalize our findings.

We introduce a framework for design solutions to address different problem areas in health portals based on the proposed improvement strategies described above.

The Refined Taxonomy of Usage Issues

Following the usage data analysis, we present the taxonomy with detailed insights in Figure 3. Note that we ground the refined taxonomy in Figure 3 on the usage data analysis. The refined taxonomy extends our knowledge of some of the existing issues and identified more specific issues, (e.g., different types of content   deficiencies). The grey text areas in Figure 3 indicate these findings, with the percentages indicating the prevalence of the issues. We empirically observed most of the issues reported in the literature.
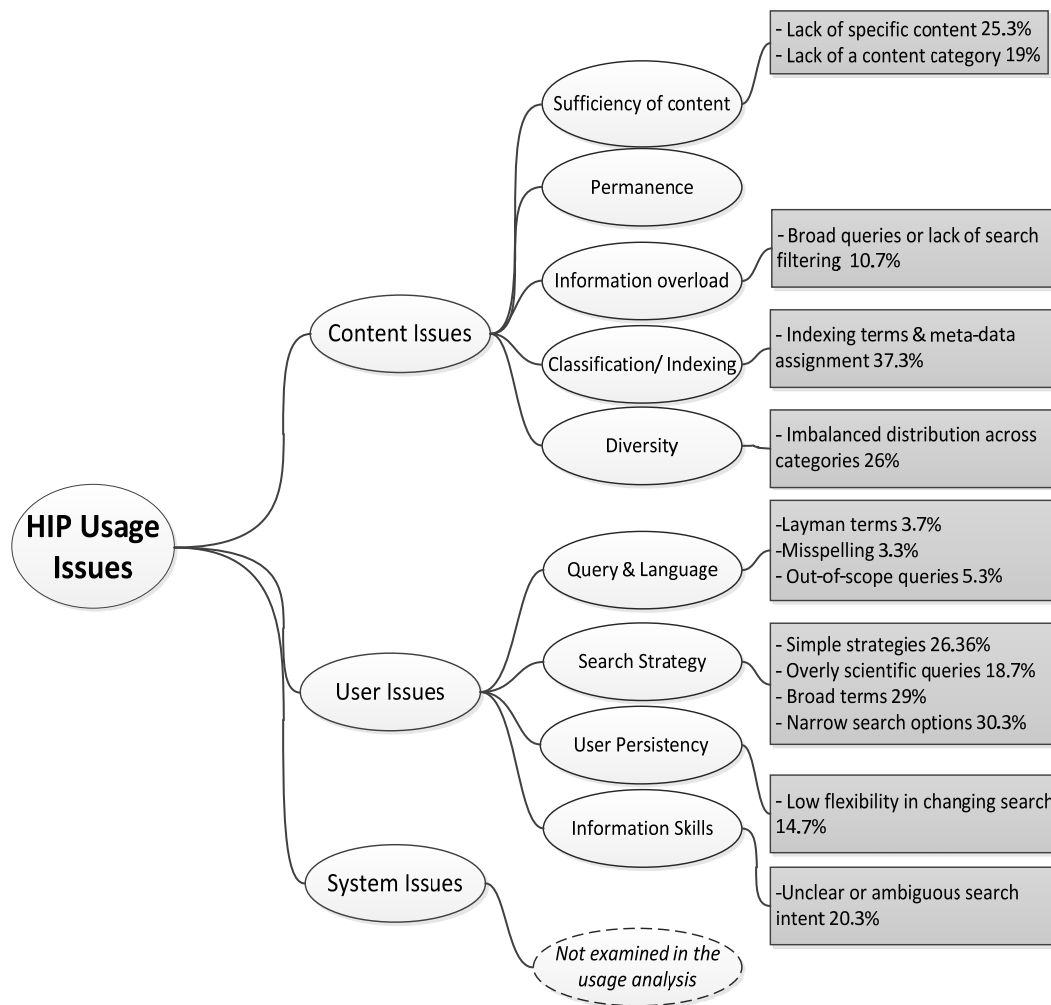

Figure 3. Refined Taxonomy of Usage Issues

4.3   Phase 3: Applying the Taxonomy of Usage Issues to Improve HIPs

Based on the insights from the usage data analysis, we investigated the role of usage data in addressing HIP issues. We demonstrate that not only can usage data be used to identify and better understand the issues, but also that it is also possible to use this data to improve the functionality of portals. Therefore, we extended the usage-driven approach (Stojanovic, Gonzalez, & Stojanovic, 2003) to improve HIPs.

Below we provide some examples on how we arrived at the recommended solutions, informed by the observation from the usage data analysis. Table 5 summarizes the solutions followed by further details.

- Solution for lack of content: we observed many cases involving a lack of content could have been avoided had there been a mechanism to alert the domain experts on legitimate searches that failed to retrieve any results. Moreover, when removing from the cases those instances

related to searching issues, the lack of content cases were more clearly identified. For instance, "breast reconstruction cost" was clearly the result of a lack of content because there were no matches in the HIP repository. We also observed that many cases might represent a missing general category in our content collection, such as information relating to cell/gene topics. Finally, we observed that it is possible to derive such insights by mining usage data with suitable information processing.

Table 5. Recommended Usage-based Actions with Examples

Issues	Recommended actions	Nature of the problems from empirical findings
Content insufficient or unbalanced (lack of diversity) C1, C2, C6	Identify missing content by reviewing failed queries that are not present in the database. In the case of C2, identify a new content category sought after by users. Revision of the portal's topic spectrum may be required.	**Recurring:** novel information needs emerge from users over time.
Lack of permanence C3	Identify expired links through analyzing "bounce-backs" in usage history.	**Recurring:** as external content changes, is updated or expires.
Irrelevant content C4	Based on usage access statistics, review unvisited content; review searches with too many results to improve searching precision.	**Recurring:** time-dependent.
Indexing problems C5	Usage-driven indexing based on real searches; acquire user-generated terms for indexing. Identify new content that better matches the profiles.	**Recurring:** as new layman terms or ways of describing the information need are employed by users.
Query & language U1,U2,U3	Identify common query faults in usage logs to improve language support tools (such as the "did you mean?" function in online dictionaries/thesauri).	**Recurring:** with respect to the source of query faults One-time: with respect to the lack of language support tools.
Search strategy U4-8	Employ usage data as a source for recommending tools (such as "others also tried this..." feature).	**One-time:** need improved recommendation tools to suggest alternative search strategies.
User effort U9, U10	Provide usage-based hints for users with low persistence.	**One-time:** need improved recommendation tools to suggest alternative search strategies.
Information skills U11,U12,U13	Employ user feedback to improve resource accessibility (such as readability or excessive content). Review resource descriptors.	**Recurring:** with respect to the perceived level of readability of the content.
UI problems S1-4	Analyze user navigation patterns to improve user interface design. Longitudinal usage analysis to validate design changes.	**One-time:** need improved design fixes.

- Solution for indexing issues: we found that there was a portal feature missing that could incorporate search terms from usage data to supplement the list of indexing terms used by the system. As a result, user-oriented terms or alternative terms, such as "pain treatment" rather than "palliative", were under-represented. We found that, if such supplementary terms could be identified from usage data, it would enrich the indexing system.

- Solutions for dealing with search strategy deficiency: we found that, in some cases, users ended a failed search without exploring alternatives. It could be that a better search strategy is not obvious to users. If users were presented with more options, such as a list of searchable terms by other users or topic browsing, it could alleviate these situations. Insights from usage data can inform such recommendations. We also suggest that, if usage data are used for recommendation, it is important to ensure that such recommendations are appropriate (e.g., fit the scope of the portal) and are guaranteed to retrieve results.

In essence, such an approach promotes learning from failed usage cases to improve the search outcomes of health consumers and to sustain the portals' resources by looking at the gaps between users' needs and the system's content and functionality. We illustrate that such a perspective is useful and relevant in designing solutions because it identifies the problematic areas in health portals that need the most attention.

Researchers and others can use the knowledge encoded in the taxonomy as a guide for systematically investigating the specialized, focused approach taken to address health information provision issues. In addition, our analytical view of usage issues suggests that we should focus not only on intrinsic content, user, or system issues but also on the issues at the interface between these dimensions.

Table 5 details our recommended actions to address the issues and includes some examples drawn from the data analysis. The table also describes the nature of the problem (whether it is a recurrent or a "one-off" problem). While the findings on recurring issues are only empirical (e.g., based on our own observations), such information is important for prioritizing changes so that more frequent occurrences are dealt with first. We also note that the notion of recurring usage issues might be a novel concept that is under-explored in the literature, which might need more work in the future.

We also note the recurring nature of some content problems in which a "one-off" fix may not be appropriate. As user needs change, a mechanism to regularly review content issues and improve the system is important for the sustainability of portals. We discuss the recurring versus "one-off" nature of the issues in the next section.

In the next section, we illustrate how the proposed functionality can be implemented using BCKOnline as the context. Specifically, we describe how the insights from usage data may be used to support these improvements and what should be extracted from the usage data to make the improvements.

4.3.1 Tools for Identifying and Reporting Usage Issues for HIP Improvements

We provide examples of the usage-data analysis tool created as an instantiation of the framework (Table 5). These examples illustrate how usage data can provide action-oriented insights to address usage issues.

A screen shot of the usage data analysis tool in Figure 4 shows an example of a failed search due to a lack of relevant content. We can classify this issue as a content issue and address it as an instance of emerging user needs. Reporting such emerging searches (searches that were not seen before or that were from a recent time period) to the HIP content management staff can help domain experts identify new information needs that are not currently covered.

In Figure 5, we can see how multiple users requested information on "reconstruction" but that the database lacked content for this particular category. This highlights an imbalance of content distribution that can be classified as a user/content gap. Similar to the case above, this issue can suggest emerging information needs and should be reported to the HIP content management team.

Figure 6 lists query terms that we can identify as faulty since they led to failed search results. For instance, some terms were misspelled and some are terms that were not used for resource indexing purposes. By looking at these search terms, the portal can identify various types of misspellings, which spellcheckers can then catch. With respect to indexing issues, it is feasible to automatically harvest consumer-oriented terms (e.g., by including those that are not currently present in the official indexing glossary). The content management team and the domain experts need to review these terms as a possible addition to the indexing glossary.
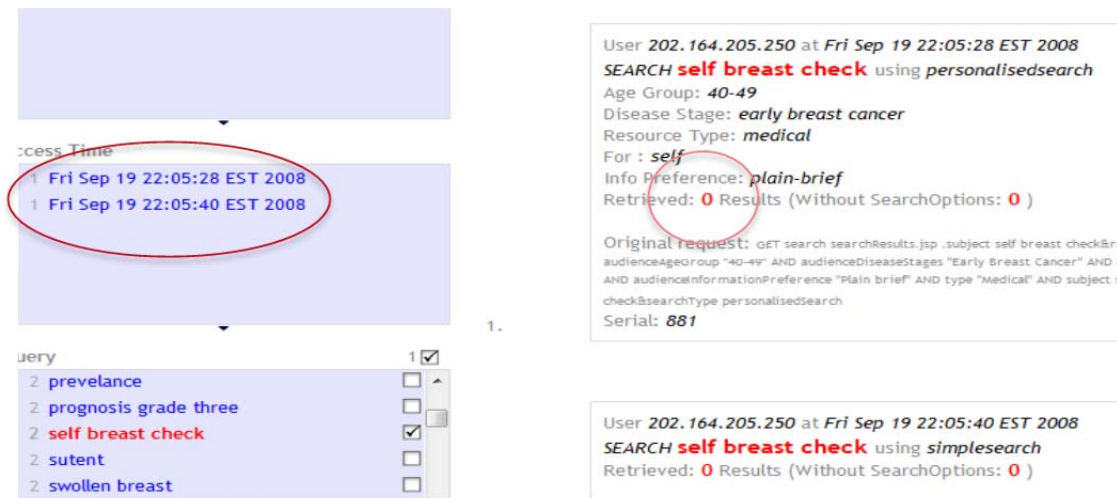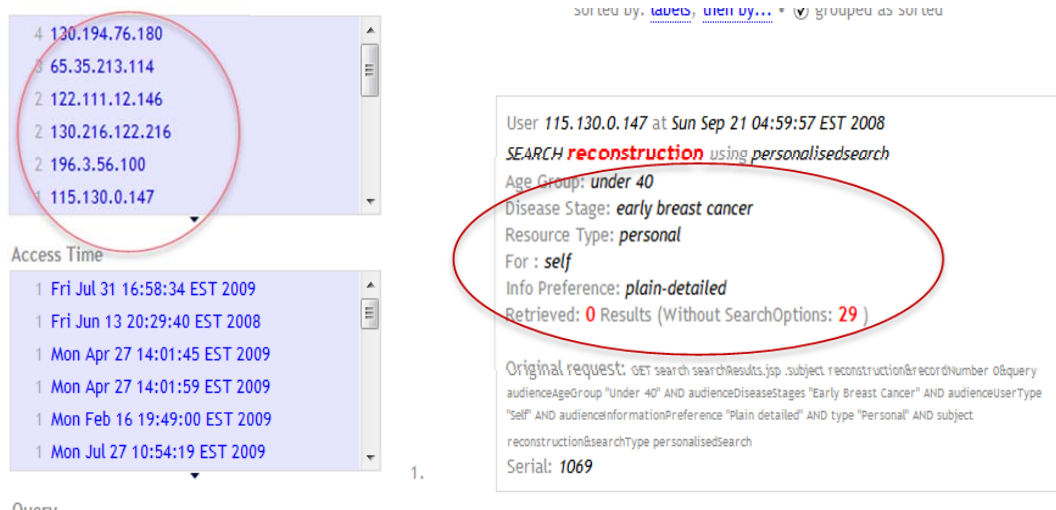
Figure 4. Lack of Content Example

Figure 5. Lack of Content or Imbalance of Content Distribution Issue
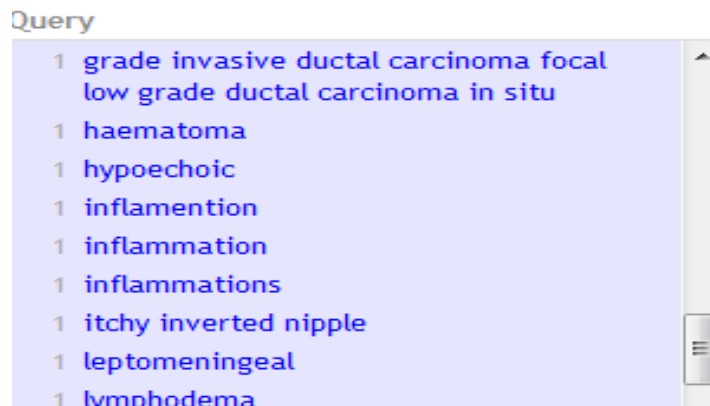

Figure 6. Sample List of Faulty Queries

These features show the potential and the feasibility of usage-driven solutions to improve health searches. However, we do not cover all of the implementation strategies, testing, and uses of these features in this paper. Longitudinal studies to measure the impacts of these changes are also an important research problem for future work.

5 Discussion

In this paper, we systematically analyze usage data for improving user searching in health portals. We developed and validated a comprehensive taxonomy of usage issues focusing on three key aspects of online health information usage: the content, the users, and the system. That is, we identify concrete examples of such issues by analyzing usage data from a health portal. We now elaborate further on the findings' significance.

5.1 The Taxonomy: Insights on Key Usage Issues

In the first phase (literature analysis), we identified key issues in health searching and provide a framework for the study. In the second phase, we obtained specific evidence of failed searches from usage data and reported the findings on content and user issues and the co-occurrence of issues.

5.1.1 Content Issues

For content issues, we found a wider range of issues than those discussed in the current literature. Most e-health papers focus extensively on lack of content (Keselman et al., 2008b), but we suggest that issues with health portals could also be due to the lack of diversity (not enough content for different "categories") or indexing issues. Note that in, this paper, we do not systematically compare the literature's findings on content issues and our findings because our literature analysis is not exhaustive.

We also identified recurring gaps between the content and usage. These gaps relate to issues of content sufficiency, indexing issues, or diversity of content. The findings highlight that some content issues are intrinsic to the content management processes in HIPs rather than to issues of external information sources. These issues point to deficiencies in the processes of content identification, revision, and content describing/indexing in HIPs.

5.1.2 User Issues

Much has been written on user issues relating to ineffective searching behavior. This study further provides specific knowledge on how these searching deficiencies occur in HIPs. Issues that we confirmed in this study include the use of lay language, query issues, ineffective search strategies and information skills.

Because users might not be aware of their search deficiencies or effective alternative strategies, different types of recommendations from the system might be needed. In the sections that follow, we discuss this matter in relation to the user support capability of HIPs.

5.1.3 System issues

Our findings on system issues, including information retrieval (e.g., search precision or relevance) and usability, is limited due to the lack of data on users' views. The key issues identified in the literature include the limitation of search support for health searching and usability and UI issues.

5.2 What we Learned about Usage Issues from Usage Data

Recent trends in data-driven research highlight the potential uses of usage data to identify issues and drive improvements . For instance, Joachims (2002, p. 1) suggests that "sufficient information is already hidden in the log files of search engines; therefore, the key problem is how to extract the relevant insights to enact system improvements. In this study, we found that several insights on usage issues can be derived from usage data. Different types of content and user issues can be identified, differentiated, and reported, which provides a framework to systematically investigate solutions to address the issues.

Our research illustrates that the gaps between content resources and user information needs can be identified from usage data. A quantitative estimate of the top co-occurring usage issues (Section 4, phase 2) also indicates problematic areas that commonly lead to failed searches, which should receive priority in

terms of improvements. Such findings echo the discussion in a white paper on consumer-centered health informatics by Keselman et al. (2008b, p. 481). These authors highlight the challenge of bridging the gaps "between the user information needs and the content of information resources". However neither study specifically identifies these gaps. This research furthers the discussion by examining what the usage gaps mean for HIPs.

In the next section, we discuss specific actions or strategies that can possibly address usage issues, in which we also illustrate the value of knowledge gained from usage data more clearly.

5.3   Using Knowledge from Usage Data to Address Usage Issues

In the third phase, built on the empirical knowledge about the utility of usage data, we recommend several solutions and strategies to deal with such issues by exploiting insights from usage data. We also theorize that such an approach could be viewed more generally as a usage-driven design strategy for health portals, which can contribute to portals' sustainability.

5.3.1   Addressing the Usage Gaps

Based on the analysis in earlier sections, we identify several intertwining issues, which are best viewed as a mismatch between content or systems' capabilities and users' skill deficiencies. This suggests that not only should we focus on the intrinsic issues in three problem categories (content/user/system), but also that it is equally important to look at failed searches that involve both the deficiencies of users and the system. We found such co-occurrences in our study.

For content/user gaps, it is desirable to have a usage-based mechanism that reports where the gaps are to effectively support the content management team and the domain experts in managing HIP content. To sustainably manage a large amount of usage data, it is also imperative to prioritize use cases that are more relevant and important, such as emerging user needs or high-profile and high-frequency failed searches.

User/system gaps are characterized by potential ineffectiveness of user search skills on the one hand and the limitation of the systems to support these situations on the other. We only observed these gaps empirically because we did not code system issues in the usage analysis. However, on replicating searches, we found several issues. In the context of HIPs, prominent issues include the lack of support for medical jargon and the use of layman terms (S3 and U1-U3), the lack of guidance for strategy recommendations (to deal with ineffective search strategies, U4 to U10), and the low tolerance in information retrieval functionality (to deal with layman, broad or narrow terms, or specialized query requests—U7, U8).

Finally, we do not here explore in detail the content/system gaps, which cover issues occurring in the translation from content management to front-end user search systems because it would require research involving the domain experts and the content management process. We leave such an inquiry to future research.

5.3.2   Supporting user Searching Issues

Norman (2002, p. 105) uses the phrase "to err is human" when discussing user errors in usability studies. Norman points out that, because people make errors routinely and many errors are often corrected automatically by the system, users are rarely aware of their own searching issues. Even though users are aware of common searching deficiencies, they still make mistakes in how they search. Given the persistent state of user issues in health searching, it sends a message] to designers that tools need to be designed to better assist the users to overcome their searching deficiencies and to provide effective alternatives to increase searching success. Clearly, learning from usage can be integrated in such solutions to provide informed, effective suggestions for users.

Implication for System Design Improvement

While we do not examine systems issues (such as information retrieval or usability) in this paper, we make several recommendations based on observing the searching deficiencies of users.

It is crucial to avoid failed searches without any system response because this can leave those searching for health information distressed, confused, and frustrated (Williamson, 2005). Failed searches with no

results are relatively prevalent given the complex searching requirement in the health domain (Hanbury, 2012) and sophisticated search tools. The usage data analysis established that empty searches and abandoned searches were prevalent in advanced search modes. However, the lack of explanation or transparency in the system's responses on the causes of as failed search could hinder users from responding and/or changing their approach. Such results highlight the need to design the search functionality of HIPs to be robust and responsive. A robust search functionality needs to be able to handle different types of searches, including ineffective queries. A responsive system needs to avoid giving users no results or no explanation on failed searches.

6   Conclusions

This research contributes to a better understanding of HIP problems based on an assembled framework of prior research and an empirical usage-data based analysis from an operational portal. While the literature has been prolific on user issues, we provide a comprehensive usage analysis across the full spectrum of usage issues in the context of health portals. We address content issues in detail, which is where the complexity of the HIP information management model translates into usage problems. These problems include lack of content, indexing issues, or issues with the diversity of the content, all of which are vital to a HIP's sustainability, as we illustrated in this paper.

The results have practical benefits for health website designers/developers and content managers. The taxonomy detailing HIP usage issues creates a mechanism for usage-based improvements for more sustainable HIPs. This taxonomy provides a clear, case-by-case understanding of HIP usage issues that one can translate directly into improvement strategies. We also propose and empirically analyze the notion of usage gaps and recurring issues and their roles in improvement strategies for health portals. We offer some strategies, based on the taxonomy, in this paper and we are investigating more. Two key strategies are as follows:

- Smart learning capability: to enable an automatic mechanism to detect problems and present the portal managers with an efficient mechanism to address content problems.
- Coping with the changing needs of users: the taxonomy highlights problems relating to the mismatch between content and user needs. A usage-based mechanism can alert the portal's manager to the existing gaps between content and users' needs by focusing on specific areas of health information (such as emerging health topics, narrow, or specific information needs).

Even though there are concerns with the reliability of usage data, the literature suggests that usage data analysis is a scalable, non-intrusive data collection method that can shed some light into a system's usage (Jansen, 2006). The vast amount of data on the Internet also means that it is an effective tool for monitoring usage in the long-term, which contributes to the longevity of health portals.

This work also has practical implications for achieving better user-centered health information provision.. There are many potential problems arising from an "information tailoring" mechanism that user-centered approaches embrace, such as problems with indexing, content selection, and personalized searches. Many of those features rely on manual processes performed by domain experts, such as content review or indexing. The taxonomy proposed in this paper provides a systematic mechanism to review and address potential problems and promotes user-centered information provision.

6.1   Limitations and Future Work

Our usage data analysis does not include underlying situational, cognitive, or affective elements of usage (Pautler et al., 2001). We did not include other dimensions of usage, such as users' reflections on their experience. Consequently, we only focused on problems that were more verifiable. While a large part of the taxonomy can be validated with usage data analysis, the results will be more reliable and perhaps more extensive if, for example, an interview method or survey was incorporated (similar to those used by Madle et al. (2006)).

The co-occurrence of issues further complicates online health provision and have not been well researched. The identification of co-occurring issues suggests a user-focused or content-focused solution alone may not be sufficient to address the problems. Based on the insights from this study, we argue that a problem-solving strategy for health portals should focus not just on the issues of each area separately, but also on the gaps between them. The identification of top co-occurring issues suggests that dealing

with these gaps needs to be prioritized. For instance, an improvement in the indexing system should focus on the use of narrow, scientific keywords. These observations may be worthy of further investigation.

With regard to the comprehensiveness of the taxonomy, future research could extend the work to other usage problems such as the information use context, language or culture, accessibility,or users' perceptions and attitudes. However, we anticipate that the central focus for research into health portals usage would still be on user/systems interaction issues and information content issues.

Finally, our taxonomy illustrates, to some extent, the potential of bridging the trend of data-driven/usage-driven research with extensive research on consumer health information seeking. Our future research agenda includes developing integrated system components that instantiate the improvement strategies outlined by the taxonomy so that we can examine their use and impact on HIPs.

Acknowledgments

References

Adams, S., & Berg, M. (2004). The nature of the Net: Constructing reliability of health information on the Web. *Information Technology & People, 17*(2), 150-170.

Alpay, L., Verhoef, J., Xie, B., Te'eni, D., & Zwetsloot-Schonk, J. (2009). Current challenge in consumer health informatics: Bridging the gap between access to information and information understanding. *Biomedical Informatics Insights, 2*(1), 1-10.

Benigeri, M., and Pluye, P. (2003) Shortcomings of health information on the Internet, *Health Promotion International*, 18, 4, 381-386.

Berland, G., Elliott, M., Morales, L., Algazy, J., Kravitz, R., Broder, M., Kanouse, D., Muñoz, J., Puyol, J., Lara, M., Watkins, K., Yang, H., & McGlynn, E. (2001). Health information on the Internet: Accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association, 285*(20), 2612-2621.

Bernstam, E. V., Herskovic, J. R., & Hersh, W. R. (2009). Query log analysis in biomedicine. *Handbook of Research on Web Log Analysis*, 359-377.

Burstein, F., Fisher, J., McKemmish, S., Manaszewicz, R., & Malhotra, P. (2005). User centred quality health information provision: benefits and challenges. In *Proceedings of the 38th Hawaii International Conference on System Sciences*.

Cline, R. J. W., & Haynes, K. M. (2001). Consumer health information seeking on the Internet: the state of the art. *Health Education Research, 16*(6), 671-692.

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research, 3*(1), 60-95.

Delone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems, 19*(4), 9-30.

Evans, J., Manaszewicz, R., & Xie, J. (2009). The role of domain expertise in smart, user-sensitive, health information portals . In *Proceedings of the 42nd Hawaii International Conference on System Sciences*.

Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the World Wide Web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ, 324*(7337), 573.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E.-R. (2002). Empirical studies accessing the quality of health information for consumers on the World Wide Web: A systematic review. *Journal of the American Medical Association, 287*(20), 2691-2700.

Find/SVP. (1998). *The HealthMed retrievers: Profiles of consumers using online health and medical information*. New York.

Fisher, J., Burstein, F., Manaszewicz, R., & Lazarenko, K. (2009). Health information portals: How can we improve the user's search experience? In *Proceedings of the 17th European Conference on Information Systems*.

Glenton, C., Paulsen, E., & Oxman, A. (2005). Portals to wonderland: Health portals lead to confusing information about the effects of health care. *BMC Medical Informatics and Decision Making, 5*(1), 7.

Hanbury, A. (2012). Medical information retrieval: an instance of domain-specific search. In *Proceedings of the 35th international ACM SIGIR conference*, Portland, Oregon, USA.

Herskovic, J., Tanaka, L. Y., Hersh, W., & Bernstam, E. V. (2007). A day in the life of PubMed: Analysis of a typical day's query log. *Journal of the American Medical Informatics Association, 14*, 212-220.

HONSurvey. (2006). "HON's Survey of Health and Medical Internet Users."   Retrieved November 2008, from http://www.hon.ch/Survey/Survey2005/raw_data.html.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research, 28*, 407-432.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining*.

Josefsson, U. (2006). *Exploring e-patents' heterogeneity: Towards personalized e-health applications*. Paper presented at the European Conference on Information Systems.

Keselman, A., Browne, A. C., & Kaufman, D. R. (2008). Consumer health information seeking as hypothesis testing. *Journal of the American Medical Informatics Association, 15*(4), 484-495.

Keselman, A., Logan, R., Smith, C. A., Leroy, G., & Zeng-treitler, Q. (2008b). Developing informatics tools and strategies for consumer-centered health communication. *Journal of the American Medical Informatics Association, 15*(4), 473-483.

Kim, K., Lustria, M., Burke, D., & Kwon, N. (2007). Predictors of cancer information overload: Findings from a national survey. *Information Research, 12*(4), 12-14.

Kogan, S., Zeng, Q., Ash, N., & Greenes, R. A. (2001). Problems and challenges in patient information retrieval: A descriptive study. In *Proceedings of the American Medical Informatics Association Symposium*.

Kwahk, J., Smith-Jackson, T. L., & Williges, R. C. (2001). From user-centered design to senior-centered design: Designing Internet health information portals*. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 45*(6), 580-584.

Luo, W., & Najdawi, M. (2004). Trust-building measures: A review of consumer health portals. *Communications of the ACM, 47*(1), 109-113.

Madle, G., Kostkova, P., Mani-Saada, J., & Roy, A. (2006). Lessons learned from evaluation of the use of the National Electronic Library of Infection. *Health Informatics Journal, 12*(2), 137-151

Marill, J. L. (2001). Designing a usable health information portal: The MedlinePlus experience from the National Library of Medicine. In N. Campbell (Ed.), *Usability assessment of library-related Websites: Methods and case studies* (pp. 100-108). Chicago: American Library Association.

McCray, A. T., & Tse, T. (2003). *Understanding search failures in consumer health information systems*. Paper presented at the AMIA Annual Symposium.

McKemmish, S., Manaszewicz, R., Burstein, F., & Fisher, J. (2009). Consumer empowerment through metadata-based information quality reporting: The breast cancer knowledge online portal. *Journal of the American Society for Information Science and Technology, 60*(9), 1792-1807.

Nguyen, B., & Burstein, F. V. (2013). Usage-driven health information portals: A concept and design. In *Proceedings of PACIS* (pp. 1-15).

Nguyen, B. V., Burstein, F., & Fisher, J. (2015). Improving service of online health information provision: a case of usage-driven design for health information portals. *Information Systems Frontiers*, *17*(3), 493-511.

Norman, D. A. (2002). *The design of everyday things.* New York: Basic Books.

O'Mahony, B. (1999). Irish health care Web sites: A review. *Irish Medical Journal, 92*(4), 334-337.

Pautler, S., Tan, J., Dugas, G., Pus, N., Ferri, M., Hardie, W., & Chin, J. (2001). Use of the Internet for self-education by patients with prostate cancer. *Urology, 57*(2), 230-233.

Pletneva, N., Cruchet, S., Simonet, M. A., Kajiwara, M., & Boyer, C. (2011). Results of the 10th HON survey on health and medical Internet use. *Studies in Health Technology and Informatics, 169*, 73-77.

Scott-Wright, A., Crowell, J., Zeng, Q., Bates, D. W., & Greenes, R. (2006). Analysis of information needs of users of MEDLINEplus, 2002-2003. In *AMIA Annual Symposium Proceedings.*

Silvestri, F. (2010). Mining query logs: Turning search usage data into knowledge. *Information Retrieval, 4*(1-2), 1-174.

Soualmia, L. F., & Darmoni, S. J. (2005). Combining different standards and different approaches for health information retrieval in a quality-controlled gateway. *International Journal of Medical Informatics, 74*(2-4), 141-150.

Stavri, Z. (2001). Personal health information-seeking: A qualitative review of the literature. In V. L. Patel, R. Rogers, & R. Haux (Eds.), *MEDINFO V.* Amsterdam: IOS Press.

Stojanovic, N., Gonzalez, J., & Stojanovic, L. (2003). ONTOLOGER: A system for usage-driven management of ontology-based information portals. In *Proceedings of the 2nd International Conference on Knowledge Capture.*

Sutcliffe, A., & Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with Computers, 10*(3), 321-351.

Tang, T. T., Hawking, D., Craswell, N., & Griffiths, K. (2004). Quality and relevance of domain-specific search: A case study in mental health. *Information Retrieval, 9*(2), 207-225.

Tang, P. C., & Lansky, D. (2005). The missing link: Bridging the patient-provider health information gap. *Health Affairs, 24*(5), 1290-1295.

Van Der Meijden, M., Tange, H., Troost, J., & Hasman, A. (2003). Determinants of success of inpatient clinical information systems: A literature review. *Journal of the American Medical Informatics Association, 10*(3), 235-243.

Williams, P., Nicholas, D., Huntington, P., & McLean, F. (2002). Surfing for health: User evaluation of a health information website. Part two: fieldwork. *Health Information & Libraries Journal, 19*(2), 98-108.

Williamson, K. (2005). Where one size does not fit all: Understanding the needs of potential users of a portal to breast cancer knowledge online. *Journal of Health Communication, 10*(6), 567-580.

Wilson, V. (2009). *Patient-centered e-health.* Hershey, PA: IGI Publications.

Yusof, M. M., Papazafeiropoulou, A., Paul, R. J., & Stergioulas, L. K. (2008). Investigating evaluation frameworks for health information systems. *International Journal of Medical Informatics, 77*(6), 377-385.

Zeng, Q., Kogan, S., & Ash, N. (2002). Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine, 41*(4), 289-298.

Zeng, Q. T., Kogan, S., Plovnick, R. M., Crowell, J., Lacroix, E. M., & Greenes, R. A. (2004). Positive attitudes and failed queries: An exploration of the conundrums of consumer health information retrieval. *International Journal of Medical Informatics, 73*(1), 45-55.

Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association, 13*(1), 24-29.

Zhang, D., Zambrowicz, C., Zhou, H., & Roderer, N. K. (2004). User information-seeking behavior in a medical Web portal environment: A preliminary study. *Journal of the American Society for Information Science and Technology, 55*(8), 670-684.

Appendix A:

Figure A1. Example of Collected Log Data

Figure A1 shows different types of data collected from the usage log, such as user ID, time, search modes, search terms, search options, search outcomes, refinement of a search. Additional information that one can view includes the history of search of the users and list of indexing terms (see also Appendix C and D). The original raw log data can be seen in the red box.

Below, we exemplify how we coded the data:

- We selected the issue from the log because it was a "failed" search (no/few results; the original search had 0 results).
- There was available content for the query (2 items of contents when removing search options) so it was not a "lack of content". Looking at the two items, we found that they were both correctly indexed (as "scientific-detailed" content type), so we recorded no indexing issue.
- For content issues: there were no content for this topic "faslodex" for "scientific-brief" type of content. We recorded a C6 (lack of content diversity). This could be due to no existing review mechanism to ensure suffiicient content for this particular category. Flagging it as an issue might alert the domain experts to pay attention to usage needs for "scientific-brief" category.
- For user issues: we noted that this query was still a highly specific query, so we coded a U6 (highly specific/scientific query). There were no language issues (layman term or misspelling). In terms of user search strategy and effort, the user did employ a variety in successive searches (looking at the history, we can see that the user changed search options and used another query term), so there were no issues. The user did not relax the search options (otherwise the user would at least have retrieved two items), so we coded a U8 (narrow search option).
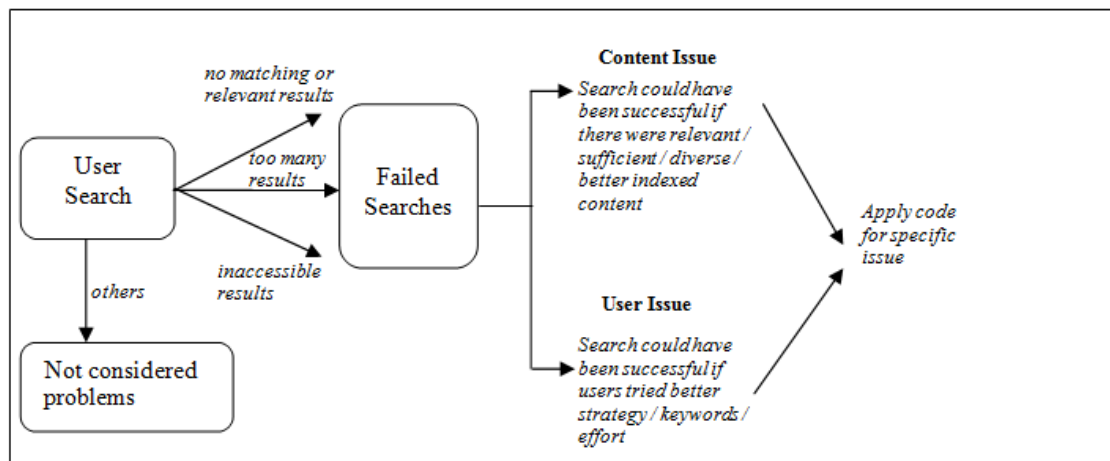- In this case, the co-occurence calculation would count C6 and U6 and C6 and U8.

Appendix B:

Figure B1. Summary of the Coding Process

www.manaraa.com

Appendix C:

Table C1. Mapping of Log Data to User Activities

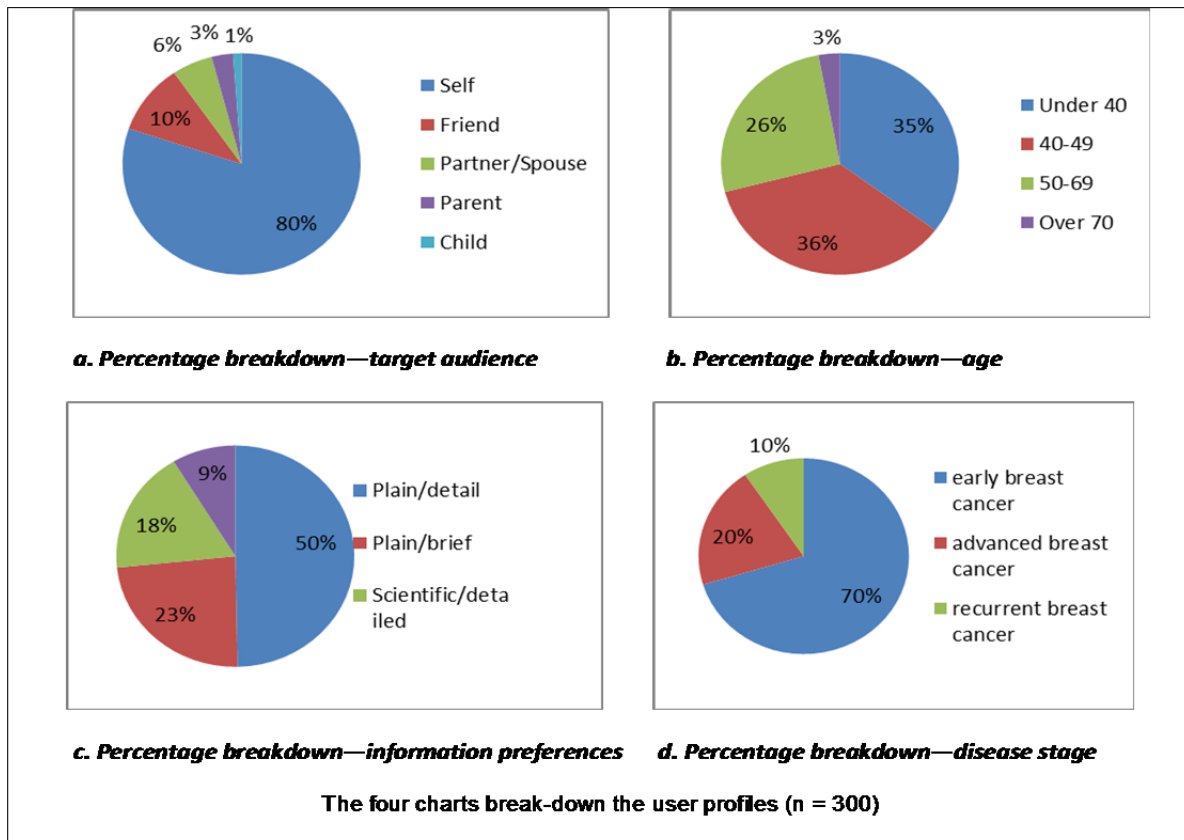| User activities	Data types	Examples
Searching (3 search modes)	User IP Search modes (personalized, topic, simple) Search terms Search options (user profiles) Time	User XXXX searched "side effects fasldoles" using "personalized.jsp", using options "disease stage: early breast cancer", "format:plain/brief"
Browsing	Search mode (Browse) Page numer	User XXXX browse to page 2 of the result pages
Obtaining results	Search terms Search outcomes	User XXXX obtained 0 results for "side effects faslodex", and 2 results when search options were removed
Refining searches	User IP (determine the same user) Search terms	User XXXX changed search term from "side effects faslodex" to "side effects" (broaden the search term)
Abandoning searches	User IP History of the same user	User XXXX did no further search after the first search

Appendix D:

a. Percentage breakdown—target audience   b. Percentage breakdown—age

c. Percentage breakdown—information preferences   d. Percentage breakdown—disease stage

The four charts break-down the user profiles (n = 300)

Figure D1. Summary Statistics of Collected Log Data

Figure D2 below provides statistics on the outcomes of searches (n = 3582). While it is clear that searches that retrieve no or few results are problematic, it is unclear whether other searches are successful. It depends on whether the documents match with the users' interests, which requires evaluation involving users. Even when a search retrieves results, it might not be successful if the results do not fit the user's need or cannot be easily consumed by users.

Note: "Others" include searches that do not fall into the above categories and were did not investigate
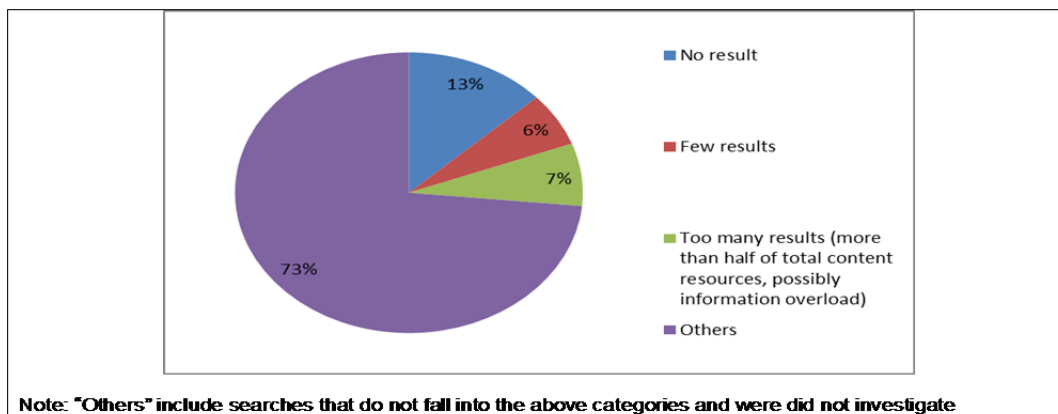
Figure D2. Outcome of Searches
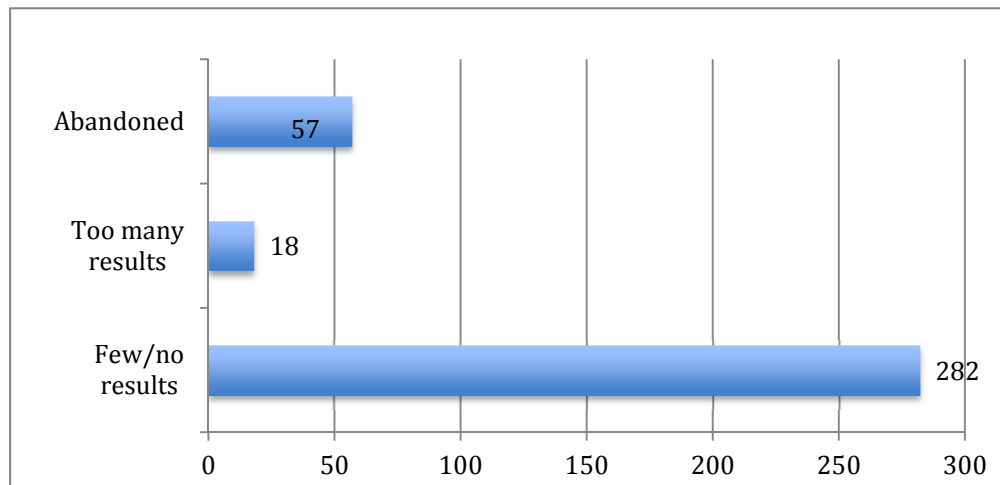
www.manaraa.com

Figure D3. Breakdown of Search Outcomes (n = 300)

About the Authors

Bang Viet Nguyen received his PhD in Information Technology from Monash University. His dissertation focused on consumer health information provision in the particular context of health information portals. Other present and past research interests include user-centered design of health systems, web information seeking and Web log analysis . He has published in conference proceedings such as those by WWW, AMCIS and PACIS.

Frada Burstein is from the Faculty of Information Technology at Monash University in Melbourne, Australia. She holds a Masters of Sci (Applied Math) from Tbilisi State University, Georgia, USSR, and a PhD in Technical Cybernetics and Information Theory from the Soviet Academy of Sciences. Her current research interests include knowledge management technologies, intelligent decision support, and health informatics. Her research appears in such journals as *Decision Support Systems*, *Journal of the American Society for Information Science and Technology*, *Journal of Organizational Computing and Electronic Commerce*, *Information Technology & People*, *Knowledge Management Research and Practice,* and others. She is a Fellow of Australian Computer Society. At Monash University, she is leading the Community Health Informatics theme for the research Flagship in ICT for Resilient Communities at the Centre for Organisational and Social Informatics (COSI).

Julie Fisher is an academic in the Faculty of Information Technology at Monash University, Australia and is currently the director of the Centre for Organisational and Social Informatics. She has worked and conducted research in the information systems field for the last 20+ years. This project involved an area of research Julie has a strong interest in, health informatics. Her other research interests include usability, particularly in relation to systems development and how to build appropriate and effective systems for users and gender and IT. She has published in numerous journals including *EJIS, Journal of the American Society for Information Science and Technology*, *Electronic Markets and Information*, and *Communication & Society*.

Campbell Wilson work also in the Faculty of Information Technology at Monash University. He holds a Bachelor of Science with Honours in applied mathematics, and Masters and PhD degrees in computer science. His research areas encompass information retrieval, machine learning and data science. He has a keen interest in the provision of high quality health information. He is also the co-director of the Faculty of Information Technology's research flagship in Data Systems and Cybersecurity.