Viewpoint

# Review of Extracting Information From the Social Web for Health Personalization

Luis Fernandez-Luque[1], MSc; Randi Karlsen[1,2], PhD; Jason Bonander[3], MA

[1]Northern Research Institute, Tromsø, Norway

[2]Computer Science Department, University of Tromsø, Tromsø, Norway

[3]Division of Knowledge Management, Centers for Disease Control and Prevention, Atlanta, GA, United States

**Corresponding Author:**
Luis Fernandez-Luque, MSc
Northern Research Institute
Postboks 6434 Forskningsparken
Tromsø, 9294
Norway
Phone: 47 93421287
Fax: 47 77629401
Email: luis.luque@norut.no

## Abstract

In recent years the Web has come into its own as a social platform where health consumers are actively creating and consuming Web content. Moreover, as the Web matures, consumers are gaining access to personalized applications adapted to their health needs and interests. The creation of personalized Web applications relies on extracted information about the users and the content to personalize. The Social Web itself provides many sources of information that can be used to extract information for personalization apart from traditional Web forms and questionnaires. This paper provides a review of different approaches for extracting information from the Social Web for health personalization. We reviewed research literature across different fields addressing the disclosure of health information in the Social Web, techniques to extract that information, and examples of personalized health applications. In addition, the paper includes a discussion of technical and socioethical challenges related to the extraction of information for health personalization.

***(J Med Internet Res 2011;13(1):e15)*** doi:10.2196/jmir.1432

**KEYWORDS**

Medical informatics; Internet, information storage and retrieval; online systems; health communication; data mining; natural language processing

## Introduction

The use of the Web by health consumers and professionals has changed with the emergence of the Social Web. This phenomenon has been described as Medicine 2.0 [1]. Whereas 10 years ago the Web was coming into its own as an e-commerce engine, the last 5 years have seen an increase in social interaction and content creation platforms that further engage and enmesh individuals in each other's online lives, increasing the sharing of knowledge. This is especially true and important for individuals seeking health information and interested in finding others with health conditions like their own. Health consumers are socializing, searching for health information [2,3], and creating content about their health in user profiles, blogs, or videos [4]. Sharing experiences and knowledge can go beyond traditional Web content and include structured health data in sites like PatientsLikeMe [5] and 23andMe [6].

The phenomenon of the Social Web would not have been possible without the transformation of Web content from static to dynamic thus providing a much richer interactive Web experience. With the emergence of the adaptive Internet in the early 1990s, websites started to change dynamically, making it possible to provide different Web content for each user. As early as 1994, the system MetaDoc changed the content of technical Web documentation based on level of expertise of the reader [7]. This adaptation of the content for a specific user is known as Web personalization [8] and adaptive hypermedia [9]. Web personalization is making the Web more efficient when accessing information and services. For example, when buying a book at Amazon.com, related recommendations are based on browsing history.

Personalization is also used to adapt Web health information and applications to the needs of each user. As explained in the background section, health education since the 1990s has been personalized and delivered through the Web with positive patient outcomes [10].

One of the main challenges when creating personalized health applications is to capture the information needed for personalization. Traditionally, information capture has relied on input from users (eg, questionnaires), which is time consuming and may undermine the interest of users. A new approach is emerging that consists of using the Web itself as a source of information for health personalization. For example, personal health records (PHRs) integrate many personalized applications, such as the online service TrialX that recommends clinical trials to health consumers based on their PHRs [11]. Content generated by health consumers can also be used for personalization. For example, in the project RiskBot, some methods have been developed for personalizing health information using data from users' profiles in MySpace [12,13]. These are just some of many examples illustrating the different possibilities for extracting information from the Social Web for health personalization.

The objective of this paper is to provide a review of the different approaches for extracting information from the Social Web for health personalization. The paper is structured as follows: the background section provides an introduction to health personalization across different research areas using as an example the case of Tailored Health Education. In the following section, we review approaches to extract information for personalization from different sources of information available in the Social Web. In the discussion section, we address current and future challenges including both technical and socioethical issues. Finally, in the conclusion we summarize the main contributions of the paper.

## Methods

In this review, our search strategies were designed to identify relevant research literature that addressed the following aspects of health personalization in the Social Web: (1) studies about the disclosure of health information in the Social Web, (2) techniques to extract that information, and (3) examples of applications. Major scientific databases in computer science (eg, ACM Digital Library) and biomedicine (eg, PubMed) were searched. In addition, we searched through the references of the selected papers, contributions to conferences, and nonresearch literature (eg, websites, books, technical reports). The background section provides an overview of the different research areas where the search was performed.

The multidisciplinary team of authors performed the selection and analysis of the relevant articles. Their backgrounds cover the different domains of the review (eg, information retrieval, computer science, health informatics, and public health). The different studies were analyzed to understand the implications for health personalization, including technical and socioethical aspects.

## Background

### Personalization

Personalization is a popular term with different meanings across domains. While personalization is the adaptation of something to a certain individual, there is a wide range of things that might be personalized (eg, treatments, websites, educational brochures, advertisements). In addition, personalization can be based on many different characteristics (eg, age, name, and location).

In the Web domain, personalization is the selection and adaptation of websites according to user specific characteristics or behaviors [8]. This is in contrast to "customization" or "adaptable systems," which refer to systems that are adapted by users themselves, for example, modifying search retrieval preferences or portal settings [9].

In medicine, the term personalization typically refers to delivering health care interventions that are designed for an individual patient (eg, drugs designed for patients with a certain genetic characteristic) [14]. However, the meaning of the word personalization varies within the health domain. In the field of tailored health education, personalization can be as simple as using the patient's name in the educational material. In that domain, personalization is a subtype of tailoring. Computer tailoring in health education has been defined as "the adaptation of health education to one specific person through a largely computerized process" [15].

For the purposes of this paper, we will use the definition of Web personalization [8] applied to the health domain. Therefore, we define Web health personalization as the adaptation of health-related Web content and applications to characteristics associated with a specific user.

### Relevant Research Areas

There are different areas of research within health informatics (see Table 1) dealing with aspects related to the acquisition of information from the Social Web for health personalization. Tailored health education, the next subsection, is of special interest because in that domain, personalized Web applications have been used for more than a decade. In addition, there are relevant research areas in computer science, which are listed in Table 2.

**Table 1.** Relevant research areas in health informatics

| Research Area | Importance for Health Personalization |
|---|---|
| Tailored health education [10] | Personalization of educational Web content to promote health and modify health behaviors |
| Personal health records [16] | PHRs are a source of information about users.<br>Personalized applications can be integrated as third party applications inside the PHRs. |
| Biomedical text mining | Data mining techniques to extract information from text, for example, automatic classification of forum posts [17] |
| Consumer health vocabulary [18] | Study of the vocabulary used by health consumers and how it maps with medical standardized vocabulary |
| Computer-aided diagnosis | Analysis of text, audio, and video for diagnosis, for example, speech analysis in neurology [19] |

**Table 2.** Relevant research areas in computer science

| Research Area | Importance for Health Personalization |
|---|---|
| User modeling and personalization | Adaptation of Web systems to users and user modeling [8] |
| Computer vision | Extraction of information from images and videos, for example, age-group classification from facial images [20] |
| Affective computing and social signaling | Extraction of information about users emotions [21] and social behavior [22] |
| Collaborative computing | Use of collaborative techniques to build personalized systems and classify content, for example, tagging of Web content [23] |
| Web data mining | Extracting information from the Web, for example, the analysis of the links to find relevant websites [24] |

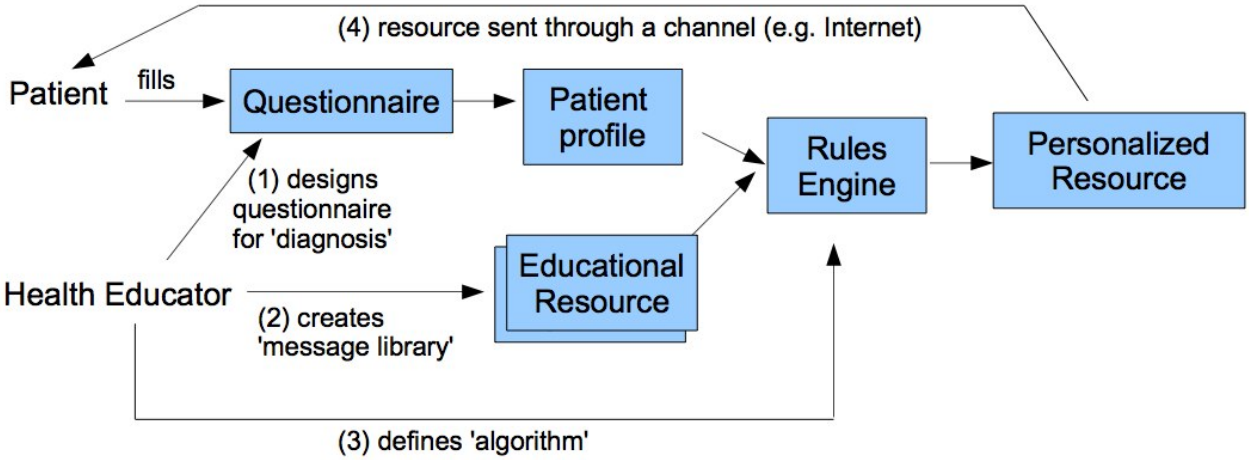
## Tailored Health Education

The origin of Web health personalization is found in the field of tailored health education. Computers have been used to personalize health education from the early 1990s, including Web educational content. Detailed reviews of personalized health education can be found in Vries et al [15], Cawsey et al [25], and Kukafka et al [26]. Reviews dealing with Web-based interventions can be found in Lustria [10], Webb et al [27], and in Enwald et al about obesity [28].

According to de Vries and Brug [15], the process of personalizing educational materials (see Figure 1) requires: at least: (1) a "diagnosis" at the individual level of characteristics that are relevant for a person's health behavior or illness; (2) a "message library" that contains all health education messages that may be needed; (3) an "algorithm," a set of decision rules that evaluates the diagnosis and selects and generates messages tailored to the specific needs of the individual user; and (4) a "channel."

Using computer science terminology, the *diagnosis* can be seen as *user modeling* and the *message library* could be seen as the repository with the Web content to personalize. Different adaptations are possible within personalized health education such as selecting which content is to be presented, ordering of content, and adaptation of content itself.

**Figure 1.** The process of tailoring health education




As described in Figure 1, most personalized health education systems can be seen as expert systems where the expertise of a human health educator is captured to create a personalized intervention (eg, text message or website) based on a set of

parameters about the patients and the educational resources. The parameters can be diverse, from basic demographics to complex psychological parameters, depending on the goals of the application. For example, physiological parameters may be more relevant to modify behaviors (eg, smoking) than to provide health information to patients with cancer. In most cases, the parameters about patients are captured using questionnaires, which are time consuming and may decrease the interest to participate. To alleviate that problem, certain parameters (eg, demographics and diagnosis) can be captured from electronic medical records [29].

Adapting the content itself can simply mean adding the patient's name in the appropriate places. It can also mean to adapt content based on behavioral parameters grounded in models such as the transtheoretical model of health behavior change [30]. Personalizing an educational brochure about smoking cessation, for example, may only add the name of the smoker to the educational materials. A more complex personalization will be to provide content with different tips depending on whether the smoker is simply contemplating quitting or has decided to quit but is worried about "side effects" (eg, gaining weight). The adaptation can also be based on demographic information such as age and gender; for example, teenagers may consider quitting smoking mainly because it damages their image (eg, yellowing teeth) and not so much because it increases the risk of cancer.

## Extracting Information From the Health Social Web

To create personalized health applications, it is necessary to acquire information about users. The information can be as simple as general demographic data (eg, age, gender, ethnicity, and location) or more complex, such as data acquired through structured questionnaires, health records, and so on. It is equally important to have adequate information about the Web content itself such as topic, language style, and date.

As summarized in Table 3, there are many information sources in the Social Web that can be used to extract information about users and content. The Social Web has facilitated the creation of Web content (eg, blogs, videos, and user profiles). User-generated content can be analyzed to extract information about Web content or users. In addition, user-generated content has been found to contain disclosed personal health information [31,32]. Further, many other sources of information are available such as ratings, links, and Web usage data (eg, click history). Finally, while not necessarily a part of the Social Web per se, personal health records (PHRs), if shared, represent a rich source of health information from which applications and services could be personalized. In the following subsections, we provide a description of different approaches to extract relevant information for health personalization from sources in the Social Web.

**Table 3.** Main sources of information for health personalization in the Social Web

| Sources | Examples of Information That Can Be Extracted for Health Personalization |
|---|---|
| Personal health records[16] | Personal health information (eg, diagnoses and treatment)<br>Demographic information<br>Genetic information (eg, rare mutations) [33] |
| Textual content | Textual content is present in most of the Web content, and it can contain information about the authors or about the content itself (eg, description of a video). |
| User profiles in online communities | Health risk behaviors (eg, smoking)<br>Demographic information [12,13,31]<br>User preferences (eg, topics of interest) [34] |
| Forum posts and comments | Personal health information (eg, diagnoses and treatments) [32]<br>Emotional/mental status of users [35]<br>Type of content (eg, informational or conversational) [36] |
| Search queries | User interests [37] |
| Tags | Topics of tagged content and users interests [38] |
| Audio | Users emotional status [39,40]<br>Diagnosis (eg, depression) [41] |
| Facial photos | Emotions [42], gender [43], and age [20] |
| Videos | Diagnosis (eg, neurological diseases) [44]<br>Characteristics of videos (eg, topic and style) [45] |
| Ratings | Users preferences and similarities [46] |
| Social networks and links | Community discovery [47,48]<br>Characteristics of Web content [24,49] |
| Web usage data | Classification of users based on navigation patterns (eg, clicks and browsing data) [50] |

### Personal Health Records

Personal health records (PHRs) are lifelong electronic sources of personal health information controlled and managed by health consumers to support decision making [16,51]. The information contained within PHRs is generated by both clinical encounters and patients themselves. Web-based PHRs are becoming increasingly available in the United States [52].

The information contained within a PHR can range from general demographics to clinical visit information, lab test results, and genetic information [16,33]. Many currently available PHRs are beginning to comply with emerging data and interoperability standards like those found with the continuity of care record (CCR), clinical document architecture (CDA) and Health Level 7's (HL7's) PHR functional model. These not only facilitate interoperability with electronic medical records (EMR) but also provide a foundation from which health applications and services can be developed.

As PHRs begin to integrate with third party applications, a larger application ecosystem is fostered, which layers additional functionality provided by the third party applications [53,54]. That approach is similar to the iTunes App Store. For example, in Microsoft HealthVault alone, there are currently upwards of 50 different third party applications [55], a good example of which is TrialX [11]. TrialX uses the data from the PHRs to find possible subjects matching the inclusion criteria in clinical trials. In the PHR Indivo, a clinical trial evaluated the use of PHRs for delivering influenza prevention education [56].

Apart from PHRs, there are patient social networking sites offering users the option to share and visualize detailed and structured personal health information within a community, for instance PatientsLikeMe [5]. However, they have yet to provide application programming interfaces (APIs) for the integration of third party applications. Some researchers are looking into the integration of PHRs with social networking [54,57].

### Textual Content

Unstructured free text is one of the most common types of generated content in the Web. As explained in the following subsections, that textual content can be from different sources: (1) user profiles, (2) forums, blogs, and comments, (3) search queries, and (4) tags.

The use of natural language processing (NLP) is the most common approach to extract information from free text. NLP is defined as the use of computer algorithms to process written and spoken human language [58]. Processing text using NLP involves several phases. It includes the extraction of keywords, stop-word removal (eg, removal of irrelevant words), word sense disambiguation and stemming (reduce words to its root). With the extracted terms, different techniques can be used to analyze them, such as terms weighting, semantic networks, and advanced data mining techniques. NLP techniques to analyze text have been enhanced with semantic technologies so that domain knowledge is taken into account in order to alleviate the ambiguity of the extracted terms [59].

Despite the scarce examples where NLP has been used to analyze health content in the Internet, it has been widely used in the biomedical domain. For instance, NLP is used to analyze biomedical text and to create information retrieval applications [60]. As a result of many years of research, several open source frameworks have been developed, such as the Unified Medical Language System (UMLS) Knowledge Source Server [61,62]. This framework provides NLP tools for analyzing biomedical text and semantic networks for matching extracted terms with standardized vocabularies.

The application of biomedical NLP for the analysis of text generated by health consumers is challenged by the gap between the medical vocabulary and the vocabulary used by the health consumers. For example, the common expression "kidney stones" may refer to the medical term *kidney calculi*. It has been found that between 20% and 50% of health consumers' expressions were not represented by professional health vocabularies [18,63]. Nevertheless, these studies imply that nearly half of the free text created by health consumers can be mapped directly to standardized medical vocabularies. Similar results have been found in self-reported symptoms of patients in PatientsLikeMe.com [64] and search queries in the MedlinePlus health portal [65]. In addition, an approach for the identification of new terms has been developed to create a consumer health vocabulary [66]. It consists of the use of NLP to find relevant terms and map them to standardized medical vocabularies. Then, the unmapped terms are classified manually and added to the consumer health vocabulary. Another possible approach to overcome the gap between the vocabularies is to recommend standardized medical terms while typing [67].

#### *User Profiles in Online Communities*

Users in social networks and online communities maintain a personal Web site with information about them. Many of these user profiles contain personal information, such as age, gender, and hobbies. Also a significant number of users disclose health information in these profiles. For example, a study found that the majority of the teenagers in MySpace are not just disclosing general demographic information but also information about their health risk behaviors (eg, alcohol abuse) [31]. In health social networks, such as TuDiabetes.com, many users disclose personal health information (eg, type of diabetes or latest blood glucose levels). A special case is PatientsLikeMe [5] where users disclose detailed health information in their profiles.

The automatic extraction of health information from profiles in social networks has been studied in the RiskBot project. In that project, NLP techniques were used to crawl, that is, explore, sex-seeking websites and classify behaviors exhibited on those sites into different risk categories with the intent of using this information to create personalized public health messages [12,13]. The same technique was recently used to extract obesity and its comorbidities from text-based hospital discharge summaries [68].

Outside the health domain, user profiles have been used to extract information about users' interests to provide recommendations and to find users with similar interests [34].

#### *Forum Posts, Blogs and Comments*

In addition to user profiles, health consumers are generating significant amounts of textual content through blogs, posts in

forums, microblogs, and comments. This content ranges from deeply personal narratives to recommendations and reviews to discrete pieces of health data. Several studies have found disclosed personal health information in different types of content (eg, Twitter [69,70] and YouTube [32]). For example, a simple search in Twitter for "#bgnow" returns tweets that include blood glucose levels. In the studies about Twitter, the extracted information was not used for personalization but was used to study the misuse of antibiotics [69] and to analyze and track sentiments, attitudes, and behavior during a pandemic [70].

Information extracted from content can also be used to gather more information about the content itself. For example, NLP techniques have been used to classify topics of health forums [17]. In this example, the posts in a medical forum were analyzed to extract terms from a predefined set of terms. Then, different data mining techniques were used to categorize the posts.

Web content can also be classified according to emotional parameters, such as intentionality, relying on the fact that the human language provides clues about emotions and intentions. The capture of these clues is being addressed in different research fields, such as affective computing [21] and opinion mining [71]. For example, a blog post can be objective and informative (eg, how to take an insulin injection) or be affective and raising a debate (eg, hate insulin injections). Techniques have already been developed outside the health domain to automatically classify posts depending on their informative nature [36]. In the health domain, similar techniques have been used to classify suicide notes [35] and preliminary work has been done in online suicide notes [72].

### *Search Queries*

Search engines are among the most popular tools to search health information [3]. Many search engines store the text entered by the users to model the previous search queries and personalize the results.

In the health domain, there are only a few examples of health search engines using search queries for personalization. These techniques are mainly used in search engines of research literature [73]. In the health portal MedlinePlus, search queries have been used to analyze the vocabulary of the health consumers [65]. However, that information is used to detect misspellings and topics of interest and not to personalize the search results.

### *Tags*

Nearly one third of Internet users in the United States have already tagged content [23] and 6% of the health information seekers have tagged or categorized Web health information [4]. Prior to the Social Web, many indexing techniques were based on taxonomies created by experts. Today, users are indexing content with their own tags that can be used collaboratively by utilizing new taxonomies of Web resources, known as "folksonomies". In addition to classifying Web content, tagging is also used to capture information about the users. For example, the tagging history of users can be used to model their interests [38].

Health-related examples of tagging are found in platforms such as TuDiabetes.com [74] and GetHealthyHarlem [75], where tags are used to search and recommend content. One of the challenges with tagging is the appearance of ambiguity between tags. The integration of tags with ontologies opens many opportunities for using semantic-enhanced techniques [76], such as giving recommendations of tags based on medical ontologies [67]. It has also been found that nearly half of the tags created by patients for describing symptoms were found in medical standardized vocabularies [64].

## Images, Video, and Audio

In the Social Web, users are creating a wide variety of content apart from the text. Video, images, and audio are gaining in popularity as vehicles for sharing experiences and opinions. Extracting information from these file types, while of interest for personalization, has its challenges. The challenges result primarily from increased interpretive ambiguity in visual and audio processing and the computational cost. While the authors are not aware of explicit projects focused on extracting information from video within the Health Social Web, there are examples in other areas of research for instance computer vision, social signaling, affective computing, and computer-aided diagnostics.

Computer vision is concerned with computer systems that extract information from images. Computer vision techniques are used in many different domains (eg, computer-aided diagnostics). There are many examples of applications that extract information from people's facial photos about emotions [42], gender [43], and age group [20].

In social signaling [22], behavioral cues (eg, vocal behavior and hand expressions) are extracted from audio, video, and pictures in order to produce a "social signal" with the meaning of the extracted information. For example, through analyzing the speech in a dialog it is possible to gather information about the emotional status of the speakers and their different roles [39,40].

Social signaling is related to affective computing [21], which aims to create systems and devices that are adapted to human emotions. Affective systems have to recognize emotional information such as the "happiness" of a video [45] or the emotional expressions in a facial photo [42].

Computer-aided diagnostics use video and audio analysis to help diagnose different pathologies. For example, voice has been used to reveal patterns in the voice of patients with depression [41] and speech alterations in neurological disorders [19]. Video has been used to quantify the tremor in patients with Parkinson [44].

## Ratings

The ability to rate content is one of the most common types of feedback in the Social Web. It is used in a wide variety of collaborative filtering applications such as recommender systems [46]. The objective of these applications is to provide personalized recommendations based on what the system knows about "you" in conjunction with what it knows about "people like you". As explained in Schafer et al [46], there are two main approaches to giving recommendations based on ratings:

item-based and user-based. Item-based recommender systems will recommend highly rated items similar to those the specific user liked before. In the case of user-based systems, the rating history of a specific user will be used to find users with similar interests. The items with highest ratings among these like-minded users will be recommended. The rationale behind item-based systems is that "people who like x also like y," while the rationale behind user-based systems is that "people similar to you also like y."

Some applications are based on ratings in the health domain. For example, the health portal HealthyHarlem integrated a rating-based recommender system of health information [77]. There are also websites with ratings of health-care providers both in the United Kingdom [78] and the United States [79]. Integration of end-user and professional ratings has been explored in the project MedCertain [80] for creating a collaborative health information filtering system.

### Social Networks and Links

In many cases, the terms "online communities" and "social networks" are used indistinguishably. However, an online community is a subtype of social network where different users interact virtually, normally sharing specific goals. A social network, in the general sense, can be any network between people, such as family networks. The study of social networks predates the Web, and it has been used in health research [81]. As explained below, social network analysis has influenced how we browse and search the Web.

Similar to human social networks, the Web is a complex network of nodes (eg, websites) that are interconnected using links. The analysis of the "linking" structure among the different websites is a common source of information about websites [24]. A link is an implicit source of information about the "authority" or "prestige" of a website. For example, an outgoing link often indicates conveyance of authority to the linked website. That principle is the basis of many Web search algorithms, such as Google's PageRank [82].

Link analysis algorithms originated from social network analysis (SNA). SNA has been used for decades as a tool to understand complex human social networks. For example, using SNA and longitudinal data from a population of people over a period of 30 years, Christakis and Fowler found important relationships between health behaviors and health risk as a product of the structure of social networks [81]. SNA has acquired more attention for the analysis of Web social networks since the Internet has become a major social platform where millions of users are establishing relationships of diverse types (eg, friends, fans, and followers).

In the domain of the Health Social Web, SNA has been used to study online communities [83]. In other Web domains, SNA has been used to extract information for personalization. For example, SNA has been used to infer characteristics (eg, centrality, reputation, and prestige) of the members of a community (eg, bloggers) [84]. That information can be used to identify nontrusted users who are more likely to have low quality ratings and content [85,86]. Another feature of SNA is the possibility to detect communities within large social networks [47,48]. The information about the subcommunities can be used for personalization. For instance, a blog about cancer from the community of forensic pathologists may not be the best to recommend to a health consumer.

Furthermore, a social network can be itself a personalization engine where users are spreading content through their friends. Individuals are using information about their friends to spread the Web content in a manual-personalized manner. This new "viral" pattern of distribution of Web content is being used in public health [87-89]. For example, the New York City Department of Health and Mental Hygiene designed an application in Facebook that let users send "e-condoms" as a mean of promoting safe sex for HIV prevention [89]. The analysis of the structure of the social network can be used to increase the dissemination of the information in viral applications by identifying users with higher influence [90].

### Web Usage Data

The extraction of Web usage data for Web personalization predates the Social Web, yet it is still widely applied. Web servers store information about users accessing websites, such as version of the Web browser, IP addresses, and clicked links. That information can be used to improve the design of a website (eg, making the most clicked elements more visible) and to personalize the interface (eg, personalizing the layout of the Web based on the size of the screen). Mobasher [50] reviews the wide range of techniques available to extract Web usage data for personalization.

Web usage data is collected in many health-related websites, such as in WebMD [91] and MedlinePlus [92]. In WebMD, Web usage data is used for personalizing the advertisements based on the type of user's Web browser. Web usage data has also been used to evaluate the impact of public health interventions [93].

## Technical and Socioethical Challenges

As explained in the previous section, there are many possible approaches to extracting information for health personalization for the Social Web. However, these approaches have different implications, and how to apply them in personalization will vary depending on the context of the application. In order to decide which approach is the most suitable for a specific application, it is necessary to take into account the main technical and socioethical challenges arising from applying these approaches in health personalization. These challenges are addressed in the following subsections.

### Technical Challenges

There is a set of technical challenges associated with the approaches addressed in the previous sections. While it is not feasible to cover all the challenges with each approach, the discussion will focus on what we consider to be the most important ones related to health personalization (Table 4).

**Table 4.** Main technical challenges of extracting information from the Health Social Web

| Challenges | Description |
|---|---|
| Relevance [94] | To determine which information is relevant for personalization is complex, and it depends on the objectives of the personalization. |
| Reliability and validity | The reliability and validity of the information used for personalizing is heterogeneous. Users can fake information about themselves [95] or the Web content they create [96]. |
| Integration | Many Health Social Web applications are not integrated. However, some platforms provide open APIs to integrate third party applications [53]. Integration across different platforms can be achieved using semantic technologies [97]. |
| Privacy-preserving extraction of personal information | Preserving privacy while user modeling and data mining [98,99] |

Technological levels of maturity vary among the different approaches reviewed in this paper. Some are not only technologically feasible, but are commonly used in health personalization (eg, using PHR data to build personalized applications). Other approaches, such as the use of social network analysis to find communities of users, are technically feasible but not yet applied in health personalization. Other approaches are still experimental or too complex to be applied, such as video analysis.

The extracted information will have different levels of reliability, and whether that information can be used will depend on the application. For example, information extracted from a user profile in MySpace may be reliable enough to target a public health intervention but hardly specific enough to personalize an intervention or find subjects for a clinical trial recommender system. In addition to the reliability of the different techniques to extract information, we have to consider the validity of the sources of information. Many users tend to fake information to protect their privacy. For example, in a study of Facebook profiles, it was found that 8% of the users had fake names [95]. A similar problem is found in Web content, where tags describing content may be fake or spam [96]. The best way to ensure reliability and validity is to have human experts evaluating them. An alternative option is to rely on several data sources. In the example of the health video, it is possible to consider the keywords provided by the author and the viewers, comments, and so on.

There are other technical challenges that are not related to the extraction of information itself, but to the different objectives of the personalization. For example, a personalized recommender system of videos for smoking cessation may suggest a video with a lung cancer x-ray. Although effective, the user may dislike and rate the video as poor. In that case, the relevance and quality of the recommendation depends on clinical parameters and not just ratings, as traditionally recommender systems do. Furthermore, different goals imply different needs of information for modeling both users and resources. A relevant parameter for a personalized application about sexual health, for example, sexual orientation, may be irrelevant in many other applications. The discussion about relevance and quality has been addressed during many years in the field of information retrieval [94,100].

In the Health Social Web, there is a wide range of data sources and applications that are not integrated. Many platforms, such as online communities, don't provide APIs for extracting information or integrating third party applications. The lack of open APIs makes it challenging to extract information for personalization and almost impossible to integrate personalized applications. However, the use of APIs is increasing as exemplified by certain PHRs that can integrate third party applications [53-55]. However, each PHR often comes with a different API, making it hard to integrate applications across different platforms. An approach to address this problem is the creation of APIs that can be used across different platforms. This approach has been applied to integrate data from different social networks platforms [97].

As explained in following subsection, one of the most important ethical challenges is how to preserve privacy while extracting information about users. That concern has motivated the creation of different data mining techniques that preserve the privacy of the "data-mined" users [98,99]. Furthermore, many Web platforms allow the users to define their own privacy preferences.

The Social Web has changed how health information and applications are being disseminated (eg, viral dissemination and collaborative filtering). Users are now relying less on traditional experts and more on guidance from fellow users within their social networks. This phenomenon, which has been termed "apomedation" [101], is already affecting personalized health applications. For example, an increasing number of applications are relying on users to be disseminated throughout their social networks [89]. This approach has implications in the evaluation of these viral applications since it may be impossible to control who uses them. One possible solution for that problem is to extract information about impact of these applications from the social network itself [93,102].

## Socioethical Challenges

While we consider ways to use available personal information to make Web content and applications more useful, we must be mindful of related ethical challenges in doing so. First and foremost among them is privacy. There is a continuum of personal information that is captured, logged, left, and made available in the Social Web. Personal health records, for example, are by definition likely to contain highly sensitive personal information and, as such, the majority of PHR providers have varied privacy and confidentiality policies as part of their terms of use. Third party applications that make use of PHR content will need to conform to stated privacy policies.

However, this will not be easy as there are no standards for PHR privacy policies. As such, it will be difficult to create a single application that could be of use across different PHRs.

Existing on the other end of the continuum are those who are intentionally disclosing personal information about themselves or loved ones within blogs (eg, blogging about family genetic risks and the health of their children) [103,104]. In these contexts, privacy and confidentiality policies rarely exist, as individuals are simply free to publicly write about whatever is on their minds. When using techniques that extract user information, it is important to maintain a proper balance between the public and private nature of the content. Researchers should be mindful about common research principles, such as informed consent for using extracted information, and may consider poststudy interventions such as those used by Moreno et al [105]. Such principles can be seen in applications that first ask users if it is appropriate to use identifiable information, such as the ability to use current location to receive "geo-located" relevant content. As Wang and Kobsa suggested, there is a need to tailor privacy to the constraints of each individual user [106]. Mayer-Schonberger, on the other hand, has argued for the important historical role "forgetting" has played in society. He extends this idea to the Web in the form of expiration dates for information [107]. This deceptively simple idea would allow the erasure of certain kinds of information from the ubiquitous and eternal memory of the Web.

Another ethical issue regarding privacy is the extracting of information about minors because they are especially vulnerable to misuses of personal information. Unfortunately, disclosure of personal health information in social networks is rather common among teenagers [31]. There are different approaches to reducing it. For example, some researchers have approached minors disclosing health information on MySpace suggesting they reduce their disclosure of sensitive information by sending them emails to their profiles [108]. These messages sent to the teenagers reduced the disclosure of personal health information, but such emails may have been seen by some teenagers as spam. To avoid the risk of being seen as spammers, one possible approach is to rely on users to disseminate the intervention through their friends.

Many personalized applications within the Social Web intend to enhance socializing and sharing of knowledge between users. Unfortunately, in the health domain, there are some scenarios where the desired goal may be the opposite, since there are online communities promoting unhealthy behaviors, such as communities promoting anorexia and bulimia as "lifestyles" [109-111]. Facilitating the sharing of "proanorexic" knowledge and socializing can be harmful. However, the approaches presented in this paper can be used to identify these communities to reduce their impact (eg, parental software filtering proanorexia communities).

The integration between different data sources in the Web is partially a technical issue, but to achieve complete interoperability, there are also other barriers to be addressed. The terms of use of many Web services and APIs are complex to understand for both users and developers. In addition, these terms are normally framed within regional or national legislation, and many users may reside in locations with different legislation. For example, consumers of a company providing online direct-to-consumer genetic services, such as 23andMe, may receive online genetic counseling, which is illegal or not regulated in many countries. In addition, the laws enforcing privacy are different in each country and this affects the development of personalized applications [112]. What can be legally extracted and stored about users changes across the different countries; thus, a personalized health application may be doing something illegal while extracting information about their users depending on their residence.

## Conclusions

The Web has largely become a social platform where millions of health consumers are accessing and sharing knowledge about health [1,4]. Health consumers are not just socializing and accessing information on the Web, but are also using an increasing number of Web applications (eg, search engines and PHRs) to improve their perceived understanding of health issues. Many of these Web health applications are personalized to each user. One key aspect of health personalization in the Social Web is to extract information about users and resources. As reviewed in this paper, the Social Web offers many possibilities for the extraction of information about users and resources. It can be as simple as extracting information about age or as complex as extracting information about emotions. These techniques can be used not only for creating personalized applications but also for public health (eg, health surveillance) as part of the emerging discipline of "infodemiology" [113].

The adaptation of online intervention methodologies [114] to the context of personalization and the Social Web is an area for further research and beyond the scope of this paper. Critical issues need further exploration such as the scope and boundaries of effective online interventions, the role of trust in online health social networks and communities, and the ethical implications of research with publicly disclosed personal health information. The development of the techniques reviewed in this paper leads to new research questions: How to use the extracted information to influence health behavior in online contexts? How can we move techniques beyond individuals to groups, communities, and populations? In addition, more research is needed to determine the intrusiveness of these techniques. We need to be mindful of the issues raised in this paper, but the challenges cannot be an excuse not to develop more dynamic and personalized health applications. Outside the health domain, Web applications are becoming increasingly personalized; thus, health consumers will expect a more personalized experience in Web health applications.

The use of different approaches reviewed in this paper can catalyze the emergence of new applications adapted to the specific needs of the users without posing the traditional burden of filling in questionnaires and forms. However, in Web personalization "one size does not fit all," so in order to decide which techniques are suitable for a specific application, we have to bear in mind the goals of the application and the personal preferences of users.

## Acknowledgments

We would like to thank the reviewers and our colleagues for their very useful comments, which helped to improve this paper. This project belongs to the Tromsø Telemedicine Laboratory cofunded by the Research Council of Norway, project 174934.

## Conflicts of Interest

None declared

## References

1. Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. J Med Internet Res 2008;10(3):e22 [FREE Full text] [doi: 10.2196/jmir.1030] [Medline: 18725354]
2. Kummervold PE, Chronaki CE, Lausen B, Prokosch HU, Rasmussen J, Santana S, et al. eHealth trends in Europe 2005-2007: a population-based survey. J Med Internet Res 2008;10(4):e42 [FREE Full text] [doi: 10.2196/jmir.1023] [Medline: 19017584]
3. Fox S. Online Health Search 2006. Washington, DC: Pew Internet & American Life Project; 2006 Oct 29. URL: http://www.pewinternet.org/Reports/2006/Online-Health-Search-2006.aspx?r=1 [accessed 2010-04-12] [WebCite Cache ID 5owi8t9ky]
4. Fox S, Jones S. The social life of health information. Washington, DC: Pew Internet & American Life Project; 2009 Jun 11. URL: http://www.pewinternet.org/~/media//Files/Reports/2009/PIP_Health_2009.pdf [accessed 2009-12-17] [WebCite Cache ID 5m5IixBMx]
5. Frost JH, Massagli MP. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. J Med Internet Res 2008;10(3):e15 [FREE Full text] [doi: 10.2196/jmir.1053] [Medline: 18504244]
6. 23andMe. URL: https://www.23andme.com/ [accessed 2009-12-17] [WebCite Cache ID 5m5Keby53]
7. Boyle C, Encarnacion AO. Metadoc: An adaptive hypertext reading system. User Modeling and User-Adapted Interaction 1994;4(1):1-19 [FREE Full text] [doi: 10.1007/BF01142355]
8. Anand SS, Mobasher B. Introduction to intelligent techniques for Web personalization. ACM Transactions on Internet Technology 2007 Oct 18;7(4) [FREE Full text] [doi: 10.1145/1278366.1278367]
9. Brusilovsky P. Methods and techniques of adaptive hypermedia. Journal of User Modeling and User-Adapted Interaction 1996;6:87-129. [doi: 10.1007/BF00143964]
10. Lustria ML, Cortese J, Noar SM, Glueckauf RL. Computer-tailored health interventions delivered over the Web: review and analysis of key components. Patient Educ Couns 2009 Feb;74(2):156-173. [doi: 10.1016/j.pec.2008.08.023] [Medline: 18947966]
11. TrialX. URL: http://trialx.com/about/ [accessed 2009-12-17] [WebCite Cache ID 5m5Ukelsc]
12. Mishra N. Online social networking: Can it be used for risk behavior surveillance. 2008 Presented at: AMIA Spring Congress; May 29-31, 2008; Phoenix, AZ URL: https://www.amia.org:443/meetings/s08/programs.asp
13. Bonander J. Personally tailored health information: a health 2.0 approach. In: Medicine 2.0 Proceedings. Toronto, Canada: Journal of Medical Internet Research; 2008 Presented at: Medicine 2.0 Congress; September 4-5, 2008; Toronto, Canada URL: http://www.medicine20congress.com
14. van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 2008 Apr 3;452(7187):564-570. [doi: 10.1038/nature06915] [Medline: 18385730]
15. de Vries H, Brug J. Computer-tailored interventions motivating people to adopt health promoting behaviours: introduction to a new approach. Patient Educ Couns 1999 Feb;36(2):99-105. [Medline: 10223015]
16. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126. [doi: 10.1197/jamia.M2025] [Medline: 16357345]
17. Himmel W, Reincke U, Michelmann HW. Text mining and natural language processing approaches for automatic categorization of lay requests to web-based expert forums. J Med Internet Res 2009;11(3):e25 [FREE Full text] [doi: 10.2196/jmir.1123] [Medline: 19632978]
18. Keselman A, Logan R, Smith CA, Leroy G, Zeng-Treitler Q. Developing informatics tools and strategies for consumer-centered health communication. J Am Med Inform Assoc 2008;15(4):473-483. [doi: 10.1197/jamia.M2744] [Medline: 18436895]
19. Peintner B, Jarrold W, Vergyriy D, Richey C, Tempini ML, Ogar J. Learning diagnostic models using speech and language measures. Conf Proc IEEE Eng Med Biol Soc 2008;2008:4648-4651. [doi: 10.1109/IEMBS.2008.4650249] [Medline: 19163752]
20. Kwon YH, Vitoria Lobo N. Age classification from facial images. Computer Vision and Image Understanding 1999 Apr;74(1):1-21. [doi: 10.1006/cviu.1997.0549]

21. Carberry S, de Rosis F. Introduction to special issue on 'affective modeling and adaptation'. Modeling and User-Adapted Interaction 2008;18(1):1-9. [doi: 10.1007/s11257-007-9044-7]
22. Vinciarelli A, Pantic M, Bourlard H, Pentland A. Social signal processing: state-of-the-art and future perspectives of an emerging domain. New York, NY: Association for Computing Machinery; 2008 Presented at: 16th ACM International Conference on Multimedia; October 26-31, 2008; Vancouver, British Columbia, Canada.
23. Rainie L. Tagging. Washington, DC: Pew Internet & American Life Project; 2007 Jan 31. URL: http://www.pewinternet.org/~/media//Files/Reports/2007/PIP_Tagging.pdf.pdf [accessed 2009-12-17] [WebCite Cache ID 5m5PleuWG]
24. Getoor L, Diehl CP. Link mining: a survey. ACM SIGKDD Explorations Newsletter 2005 Dec;7(2):3-12. [doi: 10.1145/1117454.1117456]
25. Cawsey A, Grasso F, Paris C. Adaptive information for consumers of healthcare. In: Brusilovsky P, Kobsa A, Nejdl W, editors. The Adaptive Web: Methods and Strategies of Web Personalization (Lecture Notes in Computer Science). Berlin, Germany: Springer; 2007:409-432.
26. Kukafka R. Tailored Health Communication. In: Lewis D, Eysenbach G, Kukafka R, Stavri PZ, Jimison HB, editors. Consumer Health Informatics: Informing Consumers and Improving Health Care. New York, NY: Springer; 2005:22-33.
27. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
28. Enwald HP, Huotari ML. Preventing the obesity epidemic by second generation tailored health communication: an interdisciplinary review. J Med Internet Res 2010;12(2):e24 [FREE Full text] [doi: 10.2196/jmir.1409] [Medline: 20584698]
29. Cawsey AJ, Jones RB, Pearson J. The evaluation of a personalised health information system for patients with cancer. Modeling and User-Adapted Interaction 2000;10(1):47-72. [doi: 10.1023/A:1008350913145]
30. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [Medline: 10170434]
31. Moreno MA, Parks MR, Zimmerman FJ, Brito TE, Christakis DA. Display of health risk behaviors on MySpace by adolescents: prevalence and associations. Arch Pediatr Adolesc Med 2009 Jan;163(1):27-34 [FREE Full text] [doi: 10.1001/archpediatrics.2008.528] [Medline: 19124700]
32. Fernandez-Luque L, Elahi N, Grajales FJ. An analysis of personal medical information disclosed in YouTube videos created by patients with multiple sclerosis. Stud Health Technol Inform 2009;150:292-296. [Medline: 19745316]
33. Adida B, Kohane IS. GenePING: secure, scalable management of personal genomic data. BMC Genomics 2006;7:93 [FREE Full text] [doi: 10.1186/1471-2164-7-93] [Medline: 16638151]
34. Liu H, Maes P. InterestMap: Harvesting social network profiles for recommendations. In: Proceedings of the IUI'2005 Beyond Personalization. 2005 Presented at: IUI'2005 Beyond Personalization 2005 Workshop; January 9, 2005; San Diego, CA.
35. Pestian JP, Matykiewicz P, Grupp-Phelan J, Lavanier SA, Combs J, Kowatch R. Using natural language processing to classify suicide notes. AMIA Annu Symp Proc 2008:1091. [Medline: 19006447]
36. Maxwell Harper F, Moy D, Konstan JA. Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. New York, NY: Association for Computing Machinery; 2009 Presented at: 27th International Conference on Human Factors in Computing Systems; April 4-9, 2009; Boston, MA URL: http://www-users.cs.umn.edu/~harper/publications/harper-chi2009.pdf [doi: 10.1145/1518701.1518819]
37. Qiu F, Cho J. Automatic identification of user interest for personalized search. In: WWW '06 Proceedings of the 15th International Conference on World Wide Web. New York, NY: Association for Computing Machinery; 2006 Presented at: 15th international World Wide Web Conference; May 23-26, 2006; Edinburgh, Scotland.
38. de Gemmis M, Lops P, Semeraro G, Basile P. Integrating tags in a semantic content-based recommender. In: Proceedings of the 2008 ACM Conference on Recommender Systems. New York, NY: Association for Computing Machinery; 2008 Presented at: 2nd International Conference on Recommender Systems; October 23-25, 2008; Lausanne, Switzerland. [doi: 10.1145/1454008.1454036]
39. Stoltzman W. Toward a social signaling framework: Activity and emphasis in speech [master's thesis]. Cambridge, MA: Massachusetts Institute of Technology; 2006. URL: http://dspace.mit.edu/handle/1721.1/41537 [accessed 2011-01-15] [WebCite Cache ID 5vlKfGRZA]
40. Lee CM, Narayanan SS. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 2005;13:293. [doi: 10.1109/TSA.2004.838534]
41. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. Journal of Neurolinguistics 2007;20:50. [doi: 10.1016/j.jneuroling.2006.04.001]
42. Cottrell GW, Metcalfe J. EMPATH: face, emotion, and gender recognition using holons. In: Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3. San Francisco, CA: Morgan Kaufmann Publishers; 1990 Presented at: Conference on Advances Neural information Processing Systems 3; 1990; Denver, CO URL: http://portal.acm.org/citation.cfm?id=105194

43. Moghaddam B, Yang M. Learning gender with support faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002;24(5):707-711. [doi: 10.1109/34.1000244]
44. Marino S, Sessa E, Di Lorenzo G, Lanzafame P, Scullica G, Bramanti A, et al. Quantitative analysis of pursuit ocular movements in Parkinson's disease by using a video-based eye tracking system. Eur Neurol 2007;58(4):193-197. [doi: 10.1159/000107939] [Medline: 17827965]
45. Wang J, Chng E, Xu C, Lu H, Tong X. Identify sports video shots with "happy" or "sad" emotions. 2006 Presented at: IEEE International Conference on Multimedia and Expo; July 9-12, 2006; Toronto, Canada. [doi: 10.1109/ICME.2006.262641]
46. Schafer JB, Frankowski D, Herlocker JL, Sen S. Collaborative filtering recommender systems. In: Brusilovsky P, Kobsa A, Nejdl W, editors. The Adaptive Web: Methods and Strategies of Web Personalization (Lecture Notes in Computer Science). Berlin, Germany: Springer; 2007:291-324.
47. Kumar R, Raghavan P, Rajagopalan S, Tomkins A. Trawling the Web for emerging cyber-communities. Computer Networks 1999 May 17;31:1481-1493. [doi: 10.1016/S1389-1286(99)00040-7]
48. Chin A, Chignell M. Automatic detection of cohesive subgroups within social hypertext: A heuristic approach. New Review of Hypermedia and Multimedia 2008 Jan;14(1):121-143. [doi: 10.1080/13614560802357180]
49. Agichtein E, Castillo C, Donato D, Aristides Gionis A, Mishne G. Finding high-quality content in social media. In: Proceedings of the International Conference on Web Search and Web Data Mining. New York, NY: Association for Computing Machinery; 2008 Presented at: WSDM '08; 2008; Palo Alto, CA. [doi: 10.1145/1341531.1341557]
50. Mobasher B, Cooley R, Srivastava J. Automatic personalization based on Web usage mining. Communications of the ACM 2000;43(8). [doi: 10.1145/345124.345169]
51. Steinbrook R. Personally controlled online health data--the next big thing in medical care? N Engl J Med 2008 Apr 17;358(16):1653-1656. [doi: 10.1056/NEJMp0801736] [Medline: 18420496]
52. Lake Research Partners. California HealthCare Foundation. 2010 Apr. Consumers and Health Information Technology: A National Survey URL: http://www.chcf.org/publications/2010/04/consumers-and-health-information-technology-a-national-survey [accessed 2011-01-16] [WebCite Cache ID 5vm7Qrzdk]
53. Mandl KD, Kohane IS. No small change for the health information economy. N Engl J Med 2009 Mar 26;360(13):1278-1281 [FREE Full text] [doi: 10.1056/NEJMp0900411] [Medline: 19321867]
54. Mandl KD, Kohane IS. Tectonic shifts in the health information economy. N Engl J Med 2008 Apr 17;358(16):1732-1737. [doi: 10.1056/NEJMsb0800220] [Medline: 18420506]
55. Microsoft HealthVault. URL: http://www.healthvault.com/personal/websites.aspx?type=application [accessed 2010-04-17] [WebCite Cache ID 5p3wu1MPV]
56. Bourgeois FT, Simons WW, Olson K, Brownstein JS, Mandl KD. Evaluation of influenza prevention in the workplace using a personally controlled health record: randomized controlled trial. J Med Internet Res 2008;10(1):e5 [FREE Full text] [doi: 10.2196/jmir.984] [Medline: 18343794]
57. Eysenbach G. Gunther Eysenbach's Random Research Rants. 2008 Feb 21. Google Health starts pilot test at Cleveland Clinic - and my reflections on Personal Health Records 2.0 (PHR 2.0) URL: http://gunther-eysenbach.blogspot.com/2008/02/google-health-starts-pilot-test-at.html [accessed 2011-01-16] [WebCite Cache ID 5vm9BjoVa]
58. AITopics. Natural Language URL: http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/NaturalLanguage [accessed 2009-11-30] [WebCite Cache ID 5lfqv27aq]
59. Sheth AP, Nagarajan M. Semantics-empowered social computing. IEEE Internet Computing 2009;13:76-80. [doi: 10.1109/MIC.2009.21]
60. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-144. [Medline: 18660887]
61. Browne AC, Divita G, Aronson AR, McCray AT. UMLS language and vocabulary tools. AMIA Annu Symp Proc 2003:798. [Medline: 14728303]
62. Thorn KE, Bangalore AK, Browne AC. The UMLS Knowledge Source Server: an experience in Web 2.0 technologies. AMIA Annu Symp Proc 2007:721-725. [Medline: 18693931]
63. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24-29. [doi: 10.1197/jamia.M1761] [Medline: 16221948]
64. Smith CA, Wicks PJ. PatientsLikeMe: Consumer health vocabulary as a folksonomy. AMIA Annu Symp Proc 2008:682-686. [Medline: 18999004]
65. Keselman A, Smith CA, Divita G, Kim H, Browne AC, Leroy G, et al. Consumer health concepts that do not map to the UMLS: where do they fit? J Am Med Inform Assoc 2008;15(4):496-505. [doi: 10.1197/jamia.M2599] [Medline: 18436906]
66. Zeng QT, Tse T, Divita G, Keselman A, Crowell J, Browne AC, et al. Term identification methods for consumer health vocabulary development. J Med Internet Res 2007;9(1):e4 [FREE Full text] [doi: 10.2196/jmir.9.1.e4] [Medline: 17478413]
67. Delbecque T, Jacquemart P, Zweigenbaum P. Indexing UMLS Semantic Types for Medical Question-Answering. Stud Health Technol Inform 2005;116:805-810. [Medline: 16160357]
68. Mishra NK, Cummo DM, Arnzen JJ, Bonander J. A rule-based approach for identifying obesity and its comorbidities in medical discharge summaries. J Am Med Inform Assoc 2009;16(4):576-579. [doi: 10.1197/jamia.M3086] [Medline: 19390102]

69. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: twitter and antibiotics. Am J Infect Control 2010 Apr;38(3):182-188. [doi: 10.1016/j.ajic.2009.11.004] [Medline: 20347636]
70. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS ONE 2010 Nov 29;5(11):e14118 [FREE Full text] [Medline: 21124761] [doi: 10.1371/journal.pone.0014118]
71. Pang B, Lee L. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2008;2(1):1-135. [doi: 10.1561/1500000011]
72. Huang Y, Goh T, Li Liew C. Hunting suicide notes in Web 2.0 - preliminary findings. 2007 Presented at: Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007); December 10-12, 2007; Taichung, Taiwan. [doi: 10.1109/ISM.Workshops.2007.92]
73. Chen D, Orthner HF, Sell SM. Personalized online information search and visualization. BMC Med Inform Decis Mak 2005;5:6 [FREE Full text] [doi: 10.1186/1472-6947-5-6] [Medline: 15766382]
74. tudiabetes.org. URL: http://www.tudiabetes.org/ [accessed 2009-12-17] [WebCite Cache ID 5m5PuTbKV]
75. Khan SA, McFarlane DJ, Li J, Ancker JS, Hutchinson C, Cohall A, et al. Healthy Harlem: empowering health consumers through social networking, tailoring and web 2.0 technologies. AMIA Annu Symp Proc 2007:1007. [Medline: 18694106]
76. Xu Z, Fu Y, Mao J, Su S. Towards the semantic web: Collaborative tag suggestions. In: WWW2006: Proceedings of the Collaborative Web Tagging Workshop. 2006 Presented at: The Collaborative Web Tagging Workshop; May 22, 2006; Edinburgh, Scotland URL: http://www.semanticmetadata.net/hosted/taggingws-www2006-files/13.pdf
77. Khan SA, Ancker JS, Li J, Kaufman D, Hutchinson C, Cohall A, et al. GetHealthyHarlem.org: developing a web platform for health promotion and wellness driven by and for the Harlem community. AMIA Annu Symp Proc 2009;2009:317-321. [Medline: 20351872]
78. Whitfield L. ehi ehealth INSIDER. 2008 Aug 06. Patient opinion mashes up NHS choices URL: http://www.e-health-insider.com/news/4031/patient_opinion_mashes_up_nhs_choices [accessed 2009-12-17] [WebCite Cache ID 5m5QtUuCe]
79. RateMDs.com. URL: http://www.ratemds.com/ [accessed 2009-12-17] [WebCite Cache ID 5m5QzrFht]
80. Eysenbach G, Yihune G, Lampe K, Cross P, Brickley D. Quality management, certification and rating of health information on the Net with MedCERTAIN: using a medPICS/RDF/XML metadata structure for implementing eHealth ethics and creating trust globally. J Med Internet Res 2000;2(2 Suppl):2E1 [FREE Full text] [Medline: 11720933]
81. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. N Engl J Med 2007 Jul 26;357(4):370-379 [FREE Full text] [doi: 10.1056/NEJMsa066082] [Medline: 17652652]
82. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 1998;30(1-7):107-117. [doi: 10.1016/S0169-7552(98)00110-X]
83. Takahashi Y, Uchida C, Miyaki K, Sakai M, Shimbo T, Nakayama T. Potential benefits and harms of a peer support social network service on the internet for people with depressive tendencies: qualitative content analysis and social network analysis. J Med Internet Res 2009;11(3):e29 [FREE Full text] [doi: 10.2196/jmir.1142] [Medline: 19632979]
84. Sabater J, Sierra C. Review on Computational Trust and Reputation Models. Journal Artificial Intelligence Review 2005 Sep;24(1):33-60. [doi: 10.1007/s10462-004-0041-5]
85. O'Donovan J, Smyth B. Trust in recommender systems. In: Proceedings of the 2005 International Conference on Intelligent User Interfaces. New York, NY: Association for Computing Machinery; 2005 Presented at: 2005 International Conference on Intelligent User Interfaces; January 10-13, 2005; San Diego, CA. [doi: 10.1145/1040830.1040870]
86. Massa P, Avesani P. Trust-aware recommender systems. New York, NY: Association for Computing Machinery; 2007 Presented at: ACM Recommender Systems 2007; October 19-20, 2007; Minneapolis, MN. [doi: 10.1145/1297231.1297235]
87. Gosselin P, Poitras P. Use of an internet "viral" marketing software platform in health promotion. J Med Internet Res 2008;10(4):e47 [FREE Full text] [doi: 10.2196/jmir.1127] [Medline: 19033151]
88. Wong J, Stoney P, Hawke M. Ossicular erosion by cholesteatoma: investigation by scanning electron microscopy utilizing a new preparation technique. J Otolaryngol 1991 Jun;20(3):216-221. [Medline: 1870172]
89. Chan S. The New York Times. 2009 Feb 11. City Unveils Facebook Page to Encourage Condom Use URL: http://www.nytimes.com/2009/02/12/nyregion/12econdom.html?_r=2 [accessed 2011-01-10] [WebCite Cache ID 5vdLqHahC]
90. Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. New York, NY: Association for Computing Machinery; 2003 Presented at: The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 24-27, 2003; Washington, DC. [doi: 10.1145/956750.956769]
91. WebMD. Privacy Policy URL: http://www.webmd.com/about-webmd-policies/about-privacy-policy [accessed 2010-04-12] [WebCite Cache ID 5oweTuJIO]
92. MedlinePlus. NLM Privacy Policy URL: http://www.nlm.nih.gov/medlineplus/privacy.html [accessed 2010-07-20] [WebCite Cache ID 5rMi88PKU]
93. Tian H, Brimmer DJ, Lin JM, Tumpey AJ, Reeves WC. Web usage data as a means of evaluating public health messaging and outreach. J Med Internet Res 2009 Dec;11(4):e52 [FREE Full text] [doi: 10.2196/jmir.1278] [Medline: 20026451]
94. Borlund P. The concept of relevance in IR. Journal of the American Society for Information Science and Technology 2003;54(10):913-925. [doi: 10.1002/asi.10286]

95. Gross R, Acquisti A. Information revelation and privacy in online social networks. In: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society. New York, NY: Association for Computing Machinery; 2005 Presented at: 2005 ACM workshop on Privacy in the electronic society; November 7-10, 2005; Alexandria, VA. [doi: 10.1145/1102199.1102214]
96. Koutrika G, Adjie Effendi F, Gyöngyi Z, Heymann P, Garcia-Molina H. Combating spam in tagging systems: An evaluation. ACM Transactions on the Web 2008;2:1. [doi: 10.1145/1409220.1409225]
97. Breslin GJ, Harth A, Bojars U, Decker S. Towards semantically-interlinked online communities. 2005 Jun Presented at: Second European Semantic Web Conference; May 29 to June 1, 2005; Heraklion, Crete, Greece. [doi: 10.1007/b136731]
98. Kobsa A, Schreck J. Privacy through pseudonymity in user-adaptive systems. ACM Transactions on Internet Technology 2003;3(2):149-183. [doi: 10.1145/767193.767196]
99. Kobsa A. Privacy-enhanced personalization. Communications of the ACM 2007;50:24. [doi: 10.1145/1278201.1278202]
100. Knight SA, Burn J. Informing Science Journal. 2005. Developing a framework for assessing information quality on the world wide web URL: http://inform.nu/Articles/Vol8/v8p159-172Knig.pdf [WebCite Cache ID 5m68Sl0f1]
101. Eysenbach G. From intermediation to disintermediation and apomediation: new models for consumers to access and assess the credibility of health information in the age of Web2.0. Stud Health Technol Inform 2007;129(Pt 1):162-166. [Medline: 17911699]
102. O'Grady L, Witteman H, Bender JL, Urowitz S, Wiljer D, Jadad AR. Measuring the impact of a moving target: towards a dynamic framework for evaluating collaborative adaptive interactive technologies. J Med Internet Res 2009;11(2):e20 [FREE Full text] [doi: 10.2196/jmir.1058] [Medline: 19632973]
103. Hurley M, Smith C. Patients' blogs--do doctors have anything to fear? BMJ 2007 Sep 29;335(7621):645-646. [doi: 10.1136/bmj.39343.478403.68] [Medline: 17901513]
104. Tunick R, Mednick L. Commentary: Electronic communication in the pediatric setting--dilemmas associated with patient blogs. J Pediatr Psychol 2009 Jun;34(5):585-587. [doi: 10.1093/jpepsy/jsn088] [Medline: 18701561]
105. Moreno MA, Fost NC, Christakis DA. Research ethics in the MySpace era. Pediatrics 2008 Jan;121(1):157-161 [FREE Full text] [doi: 10.1542/peds.2007-3015] [Medline: 18166570]
106. Wang Y, Kobsa A. Respecting users' individual privacy constraints in web personalization. 2007 Presented at: 11th International Conference on User Modeling; June 25-29, 2007; Corfu, Greece. [doi: 10.1007/978-3-540-73078-1_19]
107. Mayer-Schönberger V. Delete: The Virtue of Forgetting in the Digital Age. Princeton, NJ: Princeton University Press; 2009.
108. Moreno MA, Vanderstoep A, Parks MR, Zimmerman FJ, Kurth A, Christakis DA. Reducing at-risk adolescents' display of risk behavior on a social networking web site: a randomized controlled pilot intervention trial. Arch Pediatr Adolesc Med 2009 Jan;163(1):35-41 [FREE Full text] [doi: 10.1001/archpediatrics.2008.502] [Medline: 19124701]
109. Fox N, Ward K, O'Rourke A. Pro-anorexia, weight-loss drugs and the internet: an "anti-recovery" explanatory model of anorexia. Sociol Health Illn 2005 Nov;27(7):944-971. [doi: 10.1111/j.1467-9566.2005.00465.x] [Medline: 16313524]
110. Mulveen R, Hepworth J. An interpretative phenomenological analysis of participation in a pro-anorexia internet site and its relationship with disordered eating. J Health Psychol 2006 Mar;11(2):283-296. [doi: 10.1177/1359105306061187] [Medline: 16464925]
111. Norris ML, Boydell KM, Pinhas L, Katzman DK. Ana and the Internet: a review of pro-anorexia websites. Int J Eat Disord 2006 Sep;39(6):443-447. [doi: 10.1002/eat.20305] [Medline: 16721839]
112. Kobsa A. Personalized hypermedia and international privacy. Communications of the ACM 2002;45(5):64-67. [doi: 10.1145/506218.506249]
113. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
114. Murray E, Khadjesari Z, White IR, Kalaitzaki E, Godfrey C, McCambridge J, et al. Methodological challenges in online trials. J Med Internet Res 2009;11(2):e9 [FREE Full text] [doi: 10.2196/jmir.1052] [Medline: 19403465]


## Abbreviations

**API:** application programming interfaces
**CCR:** continuity of care record
**CDA:** clinical document architecture
**EMR:** electronic medical records
**HL7:** Health Level 7
**NLP:** natural language processing
**PHR:** personal health records
**SNA:** social network analysis
**UMLS:** Unified Medical Language System

*Edited by G Eysenbach; submitted 18.12.09; peer-reviewed by K Mandl, R Halkes; comments to author 20.02.10; revised version received 20.07.10; accepted 28.07.10; published 28.01.11*

*Please cite as:*
*Fernandez-Luque L, Karlsen R, Bonander J*
*Review of Extracting Information From the Social Web for Health Personalization*
*J Med Internet Res 2011;13(1):e15*
*URL: http://www.jmir.org/2011/1/e15/*
*doi:10.2196/jmir.1432*
*PMID:21278049*