Expertizer: A Tool to Assess the Expert Level of Online Health Websites

Richard ZOWALLAa,1, Martin WIESNERa, Daniel PFEIFERA

aDept. of Medical Informatics, Heilbronn University, Heilbronn, Germany

Abstract. Health-related Web sites have become a primary resource to search for information on diseases, diagnoses or treatment options. Various Web sites offer a great variety of such information. However, lay people might have difficulties to assess whether a certain article or Web site fits their individual level of understandability. Hence, they might get overwhelmed with the delivered complexity of medical information. In this paper, we present a Web browser plugin, Expertizer that supports users in order to easily assess the expert level of textual medical Web content. The plugin communicates with a Web service, which leverages pre-computed classification models based on a Support Vector Machine.

Keywords. Support Vector Machines, Natural Language Processing, Internet, Classification, Web Browser

Introduction

The Internet takes a central role in the process of health information acquisition [1]. Web users search online for health information, e.g., diseases, diagnoses and different treatments [2]. Unfortunately, it is difficult for laymen to understand online content that contains a high degree of medical expert vocabulary. Moreover, such content often refers to rather complex medical facts [3]. Therefore, it is important to supply informed patients with relevant medical information, which is not only scientifically accepted but also suits to their respective background knowledge [4].

In order to accommodate this problem, methods from the field of natural language processing and text classification can be applied. Using Support Vector Machines (SVMs [5]) it is possible to automatically classify a given medical text according to two different classes: laymen-friendly vs. expert-centric. Building on this approach, Internet users can be supported with a visual health expert level gauge. The gauge displays a medical text’s expert level as numeric value based on the text’s vocabulary ranging from totally laymen-friendly (=1) to very expert-centric (=10).

In previous work we outlined a system, which computes the medical expert level of a piece of text in a fully automated way [6]. For this paper, we present a freely available Web service and a related Web browser plugin. Together, these components give users an indication whether an entire Web page or a selected text fragment match their required degree of medical expertise. Our realization comes as a Firefox plugin and so, it seamlessly integrates into a user’s familiar browsing environment.

1 Corresponding Author; E-Mail: richard.zowalla@hs-heilbronn.de
1. Methods

1.1. Online Health Content

Finding relevant information can be a challenging task, which is especially true for health-related information. Powel et al. found at least four motivations why people use the Internet to find relevant content: “desires for reassurance, for second opinions for greater understanding of existing information and to circumvent perceived external barriers to traditional sources” [7].

However, information overload and irrelevant information are major obstacles for drawing conclusions on the personal health status and taking adequate actions [8]. Faced with a large amount of medical information on different information channels users often get lost or feel uncertain when investigating on their own.

Medical knowledge available online has increased drastically during the last decades. Many articles and an increasing amount of digital media content are available on health-related Web sites. Such health information artifacts (HIA) partly focus on health professionals (e.g., as found on PubMed); yet others are more consumer-centric (e.g., as found on Mayo Clinic [9] or WebMD [10]). Typically, the vocabulary in consumer-centric HIAs differs from expert-centric ones.

1.2. SVMs for Text Classification

SVMs originate from the field of machine learning and are known to perform well for text classification tasks [5]. The use of an SVM requires a so-called training phase [11]: During this phase the SVM is supplied with a collection of pre-processed documents, whose classes are already known (e.g., each document belongs to a single class). To do so, each document \(d\) is transformed into a document vector \(v\): \(v\) is obtained from \(d\) by computing document statistics on the frequency of terms in \(d\). For performance reasons, not all terms from \(d\) can be considered for \(v\), but only the ones from a subset \(F\) of terms taken from across all training documents. The terms in \(F\) are chosen due to their ability to best support class prediction for new documents. The process of computing \(F\) is called feature selection [12]. The result of the training phase is a so-called SVM model, which can then be applied to yet unclassified documents in order to predict their class. The general aim is to give a correct prediction in as many cases as possible.

1.3. Online Health Content Classification

Based on SVMs we had implemented a medical text classification system in previous work. In this context, we had automatically gathered HIAs from various German health content providers and trained a related SVM to distinguish between vectors of laymen-friendly and expert-suited documents in German [6].

| Table 1. Content providers used in the training phase for the English SVM. |
|-----------------|-----------------|----------------|
| Name            | Amount of Articles | Pre-Assigned Class Label |
| Pubmed          | 2,500            | Expert           |
| Medscape        | 2,500            | Expert           |
| NetDoctor       | 2,000            | Laymen           |
| WebMD           | 1,500            | Laymen           |
| MayoClinic      | 1,500            | Laymen           |
For this article we adapted those methods for English health content and obtained an SVM model for English as well. Both models are used for our browser plugin. The resulting English health content predictor is based upon a total of 10,000 training documents of well-known content providers as depicted in Table 1.

1.4. Language Detection

The previously outlined medical text classifiers rely on language-dependent SVM models. So, prior to predicting a text’s expert level, the system should automatically determine the language the text is written in [13]. This task is a text classification problem as well and thus, we address it with similar SVM techniques according to [14]. To build a related language predictor, we created a specific SVM model using n-gram features from large pre-classified English and German text collections. At runtime our system first uses the resulting language predictor in order to choose the appropriate (language-dependent) SVM model for expert level prediction. After that, the chosen SVM model for expert level prediction is applied.

1.5. Firefox Add-on SDK

Modern Web browsers provide mechanisms to extend browsing capabilities via custom plugins. The Add-on SDK of the Firefox Web browser is developed by the Mozilla Foundation and provides a set of high-level APIs, a runtime and a command-line tool for developing Firefox plugins [15]. Based on the SDK, a developer can easily implement a plugin’s user interface in HTML, CSS and establish the communication to Web service endpoints via JavaScript.

2. Results

2.1. System Architecture

Our browser plugin, Expertizer, is based on conventional HTTP-based client-server architecture with a server-side component. Using the above-mentioned SVM models, the latter first determines in which language a given piece of text is written in and then computes the expert level of that piece of (medical) text. We implemented a RESTful Web service to exchange data between the Expertizer and the server-side component.

As depicted in Figure 1, our Web service receives a URL of a given health-related Web site or a user-selected text fragment. The response containing the language and expert level prediction is represented in JSON format and is then sent back to the Firefox browser.

The aforementioned expert level ranges between 1 and 10. A low level indicates that the given text is suitable for (nearly) all readers – here laymen – whereas a level of 10 indicates a very expert-centric article, most suited for health professionals. Thus, a user is given an indication whether a health-related article might fit his or her individual level of understandability.
Figure 1: System Architecture of the Expertizer: Users access the Web service indirectly via the Expertizer plugin installed in their Firefox browser. The plugin sends REST-based HTTP-requests to the Web service endpoint, which computes the expert level for a given URL or selected text fragment. The SVM models are derived from pre-processed content sources of different health Web sites (see Section 1.3).

2.2. The Expertizer Browser Plugin

The Expertizer plugin is freely available and distributed via the plugin marketplace for Firefox [16]. The plugin integrates itself in the user’s familiar browsing environment. If a user opens the context menu within any health Web site and invokes the “Expertize site!” function, a URL is sent to the server. The result is visualized by the plugin in a coloring scheme in which a low expert level is presented in green and expert-centric content receives a red-colored indicator. Figures 2a and 2b depict the use of the Expertizer plugin for two randomly selected articles from the consumer health Web site Patient.co.uk and Pubmed, according to the methods described in [6].

Figure 2a: Expert level of a randomly selected Patient.co.uk article on “Smoking” scores an expert level of 2. The content provider explicitly states, that the content is written for lay people.

Figure 2b: Expert level of a randomly selected Pubmed publication on “Tobacco and Nicotine” scores an expert level of 10. Obviously, many expert terms are found in scientific journal articles.

Privacy is a crucial factor for user acceptance of our system. Therefore it does not capture any data, which can be used to identify an individual. I.e., no user names, no address information and no IP-address or browser fingerprints are stored on the server. However, submitted Web content (i.e. medical text) is stored for the purpose of a later scientific analysis.
3. Discussion

In this paper, we presented a browser plugin, which supports users when assessing whether a health related Web site fits their personal medical background knowledge. It is based on a Web service, which leverages support vector machine (SVM) models. Two related models are available for English and German and were trained for health related content.

For training and testing of the SVMs, we resorted to pre-classified texts according to their origin, i.e., the content provider. Due to the nature of automatic text classification, the expert level prediction cannot be guaranteed to be always correct – i.e., a few articles might rather be layman-friendly than expert-centric, or vice-versa. Nevertheless our system scores a good accuracy of about 85% [6] when performing a typical validation test from the field of automatic text classification.

As a next step, we intend to evaluate, whether the system’s ratings for medical texts are indeed comparable to humans’ assessments of those texts’ expert level. Moreover, prediction models for French or Spanish could support larger user groups.

References