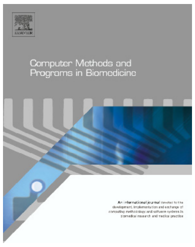
journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Automatic information timeliness assessment of diabetes web sites by evidence based medicine

Rahime Belen Sağlam[*], *Tuğba Taşkaya Temizel*

Department of Information Systems, Informatics Institute, Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 11 April 2014
Received in revised form
15 July 2014
Accepted 27 July 2014

Keywords:
Content-based retrieval
Diabetes
Medical information systems
Timeliness analysis

ABSTRACT

Studies on health domain have shown that health websites provide imperfect information and give recommendations which are not up to date with the recent literature even when their last modified dates are quite recent. In this paper, we propose a framework which assesses the timeliness of the content of health websites automatically by evidence based medicine. Our aim is to assess the accordance of website contents with the current literature and information timeliness disregarding the update time stated on the websites.

The proposed method is based on automatic term recognition, relevance feedback and information retrieval techniques in order to generate time-aware structured queries. We tested the framework on diabetes health web sites which were archived between 2006 and 2013 by Archive-it using American Diabetes Association's (ADA) guidelines. The results showed that the proposed framework achieves 65% and 77% accuracy in detecting the timeliness of the web content according to years and pre-determined time intervals respectively. Information seekers and web site owners may benefit from the proposed framework in finding relevant and up-to-date diabetes web sites.

1. Introduction

Since online content is used as an indispensable/essential source of information on health issues, the accuracy, completeness and timeliness of websites have become more of an issue for information seekers. The studies have shown that health recommendations on web sites vary in quality and many users fail to access reliable and accurate information on world wide web [1,2]. Information seekers are generally suggested to check the last update dates and the presence of any broken links in order to gain insight about the currency of a web page [1]. On the other hand, although many web sites have accurate information, recent update dates and no broken links, they may still provide outdated information [2].

In this paper, we propose a framework which automatically assesses the content of health web sites and predicts the time period that a given content belongs to. This system can serve both information seekers and website owners. Information seekers can evaluate the timeliness of health web pages easily and they can eliminate those that are not up-to-date. Website owners can utilize the proposed system to assess the currency of their web page contents automatically according to the latest research without reviewing the literature. The framework can also aid web site owners and information seekers to

[*] *Corresponding author at*: ODTÜ Enformatik Enstitüsü, Üniversiteler Mahallesi, Dumlupınar Bulvarı No. 1, 06800 Çankaya, Ankara, Turkey. Tel.: +90 533 7047710.
E-mail address: rahimebelen@gmail.com (R.B. Sağlam) .

http://dx.doi.org/10.1016/j.cmpb.2014.07.014
0169-2607/© 2014 Elsevier Ireland Ltd. All rights reserved.

identify the specifics of content that are not in line with the recent literature.

2. Related work

Information timeliness refers to information which is sufficiently up-to-date at the time of publication and it is studied under the data freshness quality dimension [3–5]. From a user's point of view, it has two sub-dimensions: currency and timeliness. Currency is estimated as the difference between the data extraction time and the data delivery time and commonly used in data warehousing systems [6]. Timeliness measures the extent to which the age of data is appropriate for the corresponding task [7].

Temporal information retrieval combines temporal relevance with document relevance and aims to return temporally relevant documents. In a well-known study, Alonso et al. [8,9] aim to extract temporal information from the documents and cluster them along a timeline supporting multiple time granularities using named-entity extraction.

A time-aware document ranking methodology relying on time-aware query suggestions is proposed by Miyanishi and Sakai [10]. Their study makes suggestions along a timeline and helps users access relevant web pages. There are also studies which date a document based on temporal language model [11]. In this approach, time partition of a document is found based on the overlapping term usage in the documents. In the study of Lin et al. [12], it is aimed to construct patient's clinical timeline from text.

There are limited number of studies which have worked on different timeliness aspects of health web sites. Post et al. manually assessed the accuracy of nutrition information on the Internet for Type 2 diabetes and concluded that the website update dates are not correlated with the accuracy of the provided information [2].

In this paper, we propose a methodology to assess the timeliness of a web site according to evidence based medicine (EBM) automatically. EBM is the conscientious, explicit and judicious use of the current best evidence in making decisions about individual patients' care by gathering the best available external clinical evidence from systematic research [13]. It aims to ensure that medical decisions are evidence based integrating both individual clinical expertise and the best external evidence. ADA guideline is the well-known gold standard for EBM on diabetes which is therefore utilized in this study.

The contributions of the paper are as follows:

1. The recent update time of a website does not necessarily indicate that its content has also been updated accordingly. The proposed methodology estimates to what time interval the content really belongs to according to a given reference guideline automatically.
2. This study is different from other existing studies in temporal information retrieval domain which focus on extraction of time related information (temporal entities) from content to predict the exact time the web page belongs to [8,14]. Such approaches are not suitable in this domain since although many dates are updated in the web pages, content may not reflect all up-to-date information in health domain. Consequently, rather than using temporal expressions, we utilize the entire document content to assess timeliness.
3. This study is different from existing studies based on temporal language models which are generally utilized for mapping events to a timeline such as [11]. Events occur at certain periods of time which has certain start and end dates. However, here, we focus on evidences about diabetes all of which have been introduced to the literature at a certain date and they become valid for a long time (majority has no end date). As a consequence, the term statistics in the corpus will be insufficient to give optimal term distribution over the time as done in [11].
4. This study is different from existing studies which aim to automate the process of determining whether a Web site is of high or low quality [15,16]. These studies do not focus on predicting the timeliness of a web site and do not take into account any temporal aspects in quality scoring.
5. The proposed methodology makes use of ADA guidelines which are gold standards for EBM in diabetes domain in order to measure the timeliness of diabetes web sites.

3. Design and methodology

3.1. *Data collection*

As there is no standard data set to test the proposed method in the literature, the data set was constructed manually by the authors. To measure the currency of web sites, ADA guidelines which were published between 2008 and 2013 are selected and retrieved from their web sites [17]. ADA guidelines consist of clinical practice recommendations specific to diabetes and are published each year based on a complete review of the relevant literature about diabetes. Several subtopics of diabetes are handled in the guidelines such as bariatric surgery, detection and diagnosis of GDM or foot care. Revisions are made each year and a brief summary of new and revisited sections is given.

The web sites to be utilized in training and testing phases are selected by querying specific search terms "diabetes" and "diabetes mellitus" in HON-search which returns the websites subscribed to the Code of Conduct (HONCode) principles [18] to ensure information quality. The HONCode is the oldest and the best-known quality label on the Web developed by a non-profit organization Health On the Net (HON). The websites which display the logo of the organization indicate that they follow the principles such as authoritativeness, statement of the purpose, confidentiality, reference section, justification of claims, website content details, disclosure of funding resources, and advertising policy.

In the data collection step, the initial HON review and subsequent monitoring dates were paid attention in order to select the corresponding archive copy of the web sites from Archive-it [19]. Archive-it is a subscription web archiving service from the Internet Archive [20] that helps organizations to harvest, build, and preserve collections of digital content. Since Archive-it does not always archive all the websites regularly, those that are missing could not be retrieved.