A web search methodology for health consumers

Marco Alfano, Biagio Lenzitti, Giosuè Lo Bosco

Abstract: Nowadays, many people use the World Wide Web to seek medical and health information but different users, such as providers (e.g., physicians) and consumers (e.g., patients), have different needs and bring different levels of reading ability and prior knowledge. Generic and specific search engines and specialized health sites either do not exploit the whole web or overload users with information. This creates difficulties mainly to consumers who often do not exactly know how to find the desired information. Thus, an information retrieval system for the web that ‘drives’ the user in finding the relevant information would be very beneficial. This paper describes a web search methodology for health consumers that help their research by classifying web pages on the basis of their level of health information and used language. We implemented such a methodology and carried out some experiments that show the effectiveness of our system when compared with the results carried out through human analysis.

Key words Consumer Health Information, Biomedical Information Retrieval, Web Search, Vocabulary.

INTRODUCTION

In the age of Internet where any kind of information can be easily found online, it is becoming increasingly evident that more and more people use the World Wide Web to seek medical and health information [3], [4], [11].

According to the 2012 Pew Internet survey on health and health care [5], 72% of Internet users say they looked online for health information within the past year (mainly for diseases and treatments). 77% of online health seekers say they began their last session at a search engine such as Google, Bing, or Yahoo. Another 13% say they began at a site that specializes in health information, like WebMD. Just 2% say they started their research at a more general site like Wikipedia and an additional 1% say they started at a social network site like Facebook.

Different users have diverse needs, even when searching for the same topic. This is certainly true in healthcare, where a patient, a physician or a health executive might look for information on the same topic but have different necessities and bring different levels of reading ability and prior knowledge together with a different vocabulary [14], [19].

Generic search engines (like Google, Bing or Yahoo) work on the whole web but make generic searches often overloading the user with the provided amount of information. Moreover, they are not able to provide specific information to different types of users [8].

On the other hand, specific search engines, such as PubMed [12] or Quertle [13], work only on medical literature (mostly PubMed). They provide extracts from medical journals that are mainly useful for medical researchers and experts but not for generic users [8]. Moreover, they do not consider all the information contained in the web that can often provide additional insights to the specific research domain being explored.
Another source of information comes from the specialized web sites oriented either to consumers (e.g., WebMD [17], Healthline [6] or MedlinePlus [10]) or professionals (e.g., Health on Net Foundation Select [7], Translating research into practice [16] or MDConsult [9]). Those sites contain very focused information but are mainly built by hand and then miss all the huge amount of information that is available on the web. Moreover, there is often a fee to be paid in order to use them.

As seen above, Internet users extensively use the web for searching health information and would greatly benefit from a search engine that provides them with the ‘right’ information they are looking for [8], [11]. This is particularly true for health customers who do not know much about the topic and the context they are exploring and can get ‘lost’ with the amount and quality of information that Internet provides [3], [4].

We have tried to put the basics for an information retrieval system of health and medical information for consumers that:

1. searches the whole web;
2. classifies web pages according to the relevance of information being searched;
3. provides web pages which use a language that can be easily understood by consumers.

The paper is organized as follows. The second chapter describes the basic principles of a web search methodology that allows health consumers to find the most suitable web pages. The third chapter describes the implementation details of our methodology and some experimental results. The final chapter presents some conclusions and future work.

SEEKING CONSUMER HEALTH INFORMATION ON THE WEB

‘Health seekers’ can be defined as Internet users who search for online information on health topics, whether they are acting as ‘health consumers’ or ‘health providers’. In particular, ‘health consumers’ can be broadly defined as patients, patients’ friends/relatives, and citizens in general [3], [4].

As said above, consumers and providers will usually look for online information with different needs, prior knowledge, vocabulary and reading skills. The vocabulary that health consumers use to discuss health topics will be the same they use to understand the content of health-related web pages. Moreover, during an online search, consumers will tend to ‘fill in’ gaps in comprehension (correctly or incorrectly) using their own knowledge, experience and preferences, since the consumer language shares some terms and concepts with the professional language [18], [19].

When searching for health information online, through for example a search engine, it would be important for consumers to receive suitable information in terms of web pages that bring information relevant to the health topic domain being explored and use a language that can be easily understood by them [11]. Thus, firstly there should be a mechanism that recognizes web pages related to the health topic searched by the user. Think, for example, of words such as heart, fever or cold that can be related to health but can also be used in different contexts. A generic search engine, like Google, will not be able to understand that the user is looking for health information and will provide all sorts of links related to those words. It will then be up to the user to visually recognize relevant web pages to him/her and this can be time-consuming. A mechanism that automatically separates “relevant” pages from “non-relevant” ones will surely help users in this task.
The other fundamental aspect is to provide consumers with web pages that use a language familiar to them and can be easily understood [15]. Also in this case, a generic search engine is not able to select pages on the basis of the type of user making the request. The consequence is that a search engine, on the same topic, will mix pages with generic information and simple language with pages with specific information and technical (medical) language.

Over the years, researchers have found that consumer terms are not well covered by the existing health vocabularies, which mostly represent the language of health professionals [18]. Indeed, expressions used by consumers to describe health-related concepts and relationships among such concepts frequently differ on multiple levels (i.e., syntactic, conceptual and explanatory) from those of professionals. As a consequence, consumer health vocabularies (CHVs) have been created for translating medical terms and concepts in their equivalent for consumers and they can be very useful for understanding if a health text is suitable to consumers or not [19].

To this end, we consider the “Open Access Collaboratory Consumer Health Vocabulary (OAC-CHV)” created and maintained by the “Consumer Health Vocabulary Initiative”\(^1\). It is a relationship file that links commonly used words to associated medical terminology represented by the “Unified Medical Language System (UMLS)\(^2\), a connection of different medical vocabularies. Differently from UMLS (and most of its source vocabularies), the OAC-CHV focuses on expressions and concepts that are employed by health-related communications from or to consumers.

The OAC-CHV contains around one hundred and sixty thousand rows (one for each term) and different fields among which:

- **CHV Term**: the term as found in text;
- **CHV Preferred Name**: the preferred consumer term as defined in the Consumer Health Vocabulary;
- **UMLS Preferred Name**: the preferred ‘medical’ term as defined by UMLS.

If we take a document (e.g., a web page) and extract the text contained in it, we can analyze the text through the OAC-CHV vocabulary to verify whether the words of the text (as single words or combination of words) are contained in it. We can then consider that a document is interesting for the consumer if it has a high number of elements belonging to the OAC-CHV vocabulary in terms of CHV preferred names (because these are the terms that consumers prefer when dealing with health topics) and also a relatively high number of UMLS preferred names (because it is important that some medical terms appear in the text in order to be sure that the document deals with health topics and because, as seen above, consumers share some terms with professionals).

When applying this process to a web search, we can analyze the results provided by the search engine and indicate, for each web page, if it has a medical content or not and if it is more oriented to the consumer or the professional. The user will then be able to concentrate on the pages that are the most interesting for his/her objectives.

---

IMPLEMENTATION AND EXPERIMENTAL RESULTS

We have implemented the methodology presented above. The system takes one or more health keywords as input and classifies web pages in a ‘consumer-oriented’ way so to provide consumers with the most suitable ones. Fig. 1 shows its basic architecture.

![System architecture diagram]

Figure 1. System architecture.

The QUERY module takes the health keyword(s) specified and the number n of web pages to be analyzed. It then searches this keyword(s) through Google and takes the first n results creating a collection of n pages with related links. For each link, the CHV ANALYZER module retrieves the related web page, cleans it, by removing tags and stop (common) words and then verifies whether each extracted term (word or combination of two or more consecutive words, e.g., ‘diabetes mellitus’) is contained in the “CHV Term” field of the OAC-CHV vocabulary. In this case, the analyzer stores the extracted term together with the number of its occurrences in the DB database incrementing, at the same time, the Terms counter, otherwise it disregards the term and increments the Non_terms counter. Upon finishing the analysis of a page, the sum of the Terms counter and the Non_terms counter will provide the total number of terms, Tot_terms, of that page. Moreover, the CHV ANALYZER verifes if each extracted term is a unique term incrementing, in this case, the Unique_terms counter, if it is a “CHV Preferred Name” (as described above) incrementing the related counter (CHV_pref_name), and if it is a “UMLS Preferred Name” (as described above) incrementing the related counter (UMLS_pref_name) only if the term is not also a “CHV Preferred Name”.

For each page, the AUTOMATIC CLASSIFIER takes into consideration the following descriptors:

\[ D_1 = \frac{CHV\_pref\_name \times UMLS\_pref\_name}{Tot\_terms} \]
\[ D_2 = \frac{Unique\_terms \times UMLS\_pref\_name}{Tot\_terms} \]

We have chosen these descriptors because they are the ones that provided the best results when executing some preliminary experiments. Nevertheless, there is also an explanation for them. \(D_1\) is used because we want to find pages with a ‘good’ medical content and this is obtained when the page presents a ‘high’ number of medical terms for both consumers (CHV_pref_name) and providers (UMLS_pref_name). Moreover, we divide by the total number of terms in order to avoid advantaging pages with a lot of text. \(D_2\) is used because we want to separate pages with a strong ‘technical’ content from pages with a more ‘general’ content. This is obtained when the page presents a high number of unique technical terms (Unique_terms) that is a characteristic of ‘specialized’ pages and a ‘high’ number of medical terms for providers (UMLS_pref_name). Moreover,
as before, we divide by the total number of terms in order to avoid advantaging pages with a lot of text.

The classification is made up of two phases:

Phase 1. The classification rule is the following:

\[
\text{If } D_1 \geq \text{Avg}(D_1) \quad \text{then the page is put into the HIGH class}
\]  
\[
\text{else the page is put into the LOW class}
\]

where:
- \( \text{Avg}(D_1) \) is the average of \( D_1 \) in the page collection;
- the HIGH class contains the most meaningful web pages (high health content);
- the LOW class contains the least meaningful pages (low health content).

Phase 2. The classification rule is the following:

For the HIGH class:

\[
\text{If } D_2 \geq \text{Avg}(D_2) \quad \text{then the page is put into the DARK GREEN class}
\]  
\[
\text{else the page is put into the LIGHT GREEN class}
\]

For the LOW class:

\[
\text{If } D_2 \geq \text{Avg}(D_2) \quad \text{then the page is put into the YELLOW class}
\]  
\[
\text{else the page is put into the RED class}
\]

where:
- \( \text{Avg}(D_2) \) is the average of \( D_2 \) in the page collection;
- The DARK GREEN class contains the pages presenting a high amount of health content related to the searched term (such as a page that describes a symptom or a disease) and are mainly written in a technical (medical) language;
- The LIGHT GREEN class contains the pages presenting a high amount of health content related to the searched term (such as a page that describes a symptom or a disease) and is mainly written in a simple (consumer) language.
- The YELLOW class contains the pages presenting a low amount of health content related to the searched term (such as a page of an institute that deals with a disease);
- The RED class contains the pages presenting no health content related to the searched term (such as a page that describes the term in another context, e.g., basketball fever).

To validate our methodology we ran some experiments assuming that users search for health terms (diseases, symptoms, treatments, etc.) in order to understand and learn more on them. We used ten health terms and searched them through Google. In particular, five terms were chosen from the list of the most searched health terms in Google (in 2013)\(^3\), i.e., \textit{flu}, \textit{cold}, \textit{labour}, \textit{balance} and \textit{diet}. Two terms were generic words that are used in different contexts, including health, such as \textit{heart} and \textit{fever}. Two terms were the most searched diseases in Google (may 2013)\(^4\), i.e., \textit{cancer} and \textit{diabetes}. As last term, we chose the medical term for diabetes, i.e., \textit{diabetes mellitus}, to evaluate the potential result differences.

\(^3\) http://www.medicalpracticeinsider.com/blog/google-zeitgeist-most-searched-health-issues-and-symptoms-2013  
\(^4\) http://www.pharmaforward.com/wp-content/uploads/2013/05/top50conditions_may2013.pdf
We searched each term through Google and retrieved the first fifty results, that is often the maximum number of results examined by a generic user. On those pages, we carried out a very basic (visual) analysis classifying each page with the four-level classification (DARK GREEN, LIGHT GREEN, YELLOW and RED) described above. This human classification constituted our ground truth for evaluating the performance of the automatic classifier.

As a first step, we used our system to analyze the pages and split them into the HIGH and LOW classes (Phase 1). Table 1 shows, for each term, the number of pages ‘correctly’ classified in each class compared to the pages visually classified. The last column contains the total number of “correctly-classified” pages. On average, 76% of pages were correctly put in the HIGH class, 82% of pages were correctly put in the LOW class and 78% of pages were overall correctly classified.

Table 1. No. of pages correctly included into the HIGH and LOW classes.

<table>
<thead>
<tr>
<th>Term</th>
<th>HIGH</th>
<th>LOW</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>10/19</td>
<td>28/31</td>
<td>38/50</td>
</tr>
<tr>
<td>cold</td>
<td>8/10</td>
<td>32/40</td>
<td>40/50</td>
</tr>
<tr>
<td>labor</td>
<td>1/1</td>
<td>34/49</td>
<td>35/50</td>
</tr>
<tr>
<td>balance</td>
<td>2/2</td>
<td>34/48</td>
<td>34/50</td>
</tr>
<tr>
<td>diet</td>
<td>10/14</td>
<td>30/36</td>
<td>40/50</td>
</tr>
<tr>
<td>heart</td>
<td>6/9</td>
<td>34/41</td>
<td>40/50</td>
</tr>
<tr>
<td>fever</td>
<td>12/17</td>
<td>31/33</td>
<td>43/50</td>
</tr>
<tr>
<td>cancer</td>
<td>9/14</td>
<td>33/36</td>
<td>42/50</td>
</tr>
<tr>
<td>diabetes</td>
<td>12/21</td>
<td>25/29</td>
<td>37/50</td>
</tr>
<tr>
<td>diabetes mellitus</td>
<td>18/29</td>
<td>18/21</td>
<td>36/50</td>
</tr>
</tbody>
</table>

Note that, on average, 30% of pages were put into the HIGH class and 70% pages were put into the LOW class, thus allowing a user to focus on a small amount of results.

As a second step we used our system to analyze the pages and put them into the DARK GREEN, LIGHT GREEN, YELLOW and RED classes (Phase 2). Table 2 shows, for the same terms, the number of pages ‘correctly’ classified in each class compared to the number of pages visually classified. The last column contains the total number of “correctly-classified” pages. On average, 44% of pages were correctly put in the DARK GREEN class, 52% of pages were correctly put in the LIGHT GREEN class, 45% of pages were correctly put in the YELLOW class, 54% of pages were correctly put in the RED class and 50% of pages were overall correctly classified.

Table 2. No. of pages correctly classified as DARK GREEN, LIGHT GREEN, YELLOW and RED.

<table>
<thead>
<tr>
<th>Term</th>
<th>DARK GREEN</th>
<th>LIGHT GREEN</th>
<th>YELLOW</th>
<th>RED</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>flu</td>
<td>6/14 (43%)</td>
<td>1/5 (20%)</td>
<td>9/21 (43%)</td>
<td>7/10 (70%)</td>
<td>23/50 (46%)</td>
</tr>
<tr>
<td>cold</td>
<td>3/9 (33%)</td>
<td>1/1 (100%)</td>
<td>2/3 (66%)</td>
<td>21/37 (57%)</td>
<td>27/50 (54%)</td>
</tr>
<tr>
<td>labor</td>
<td>0/1 (0%)</td>
<td>0/0</td>
<td>0/2 (0%)</td>
<td>16/47 (34%)</td>
<td>16/50 (32%)</td>
</tr>
<tr>
<td>balance</td>
<td>0/1 (0%)</td>
<td>1/1 (100%)</td>
<td>1/2 (50%)</td>
<td>19/46 (41%)</td>
<td>21/50 (42%)</td>
</tr>
<tr>
<td>diet</td>
<td>6/11 (55%)</td>
<td>2/3 (66%)</td>
<td>5/16 (31%)</td>
<td>14/20 (70%)</td>
<td>27/50 (54%)</td>
</tr>
<tr>
<td>heart</td>
<td>2/4 (50%)</td>
<td>3/5 (60%)</td>
<td>9/20 (45%)</td>
<td>12/21 (57%)</td>
<td>25/50 (50%)</td>
</tr>
<tr>
<td>fever</td>
<td>7/13 (54%)</td>
<td>3/4 (75%)</td>
<td>4/6 (67%)</td>
<td>19/27 (70%)</td>
<td>33/50 (66%)</td>
</tr>
<tr>
<td>cancer</td>
<td>4/6 (66%)</td>
<td>3/8 (38%)</td>
<td>17/33 (52%)</td>
<td>3/3 (100%)</td>
<td>27/50 (54%)</td>
</tr>
<tr>
<td>diabetes</td>
<td>7/17 (41%)</td>
<td>1/4 (25%)</td>
<td>9/20 (45%)</td>
<td>7/9 (83%)</td>
<td>24/50 (48%)</td>
</tr>
<tr>
<td>diabetes mellitus</td>
<td>3/8 (38%)</td>
<td>12/21 (57%)</td>
<td>4/10 (40%)</td>
<td>7/11 (64%)</td>
<td>26/50 (52%)</td>
</tr>
</tbody>
</table>
The four-level classification provides the user with more indications on the page and is potentially more useful to the consumer in helping him/her to find the relevant information but it is less precise in terms of accuracy although, as seen into the two-level classification, most of the errors happen in exchanging DARK GREEN pages with LIGHT GREEN ones and vice versa, and YELLOW pages with RED ones and vice versa, thus remaining inside the HIGH and LOW classes.

To show a practical use of our methodology, Fig. 2 presents the first ten Google results for the flu term together with their automatic classification (all the experimental results can be found at the address http://cs.unipa.it/H_search/). In this case, for example, the user could concentrate on three out of ten results, considering the LIGHT GREEN pages, or four out of ten results, considering the LIGHT GREEN and DARK GREEN pages. Of course, our system only provides an indication to the user who remains fully in charge on what to analyze. From this point of view, the potentially incorrect classification of pages does not produce particular consequences to the user search process.

CONCLUSIONS AND FUTURE WORK

This paper has presented a methodology that automatically classifies web pages so to provide health customers with the most meaningful and understandable ones. In particular, our methodology allows to split the web pages in either two classes, HIGH (high medical content) and LOW (low medical content), or four classes, DARK GREEN (high medical-expert content), LIGHT GREEN (high medical-consumer content), YELLOW (low medical content) and RED (no medical content).

The experimental results are encouraging and show the effectiveness of our system when compared with the results carried out through a human analysis. Of course, the methodology needs to be refined (e.g., considering other vocabularies) and more experiments are necessary to increase the precision of the automatic classifier, mainly for the four-level classification. The classification through human analysis must also be optimized in order to have a more objective base of comparison.

Further analyses on the web pages can provide the user with specific indicators on the page, such as main and secondary topics, complexity level and multimedia level [1], so that the user has a good knowledge on the page before analyzing it. This can also be very useful to help the user in his/her navigational path in terms of understanding the
searched topic, going into the details (deepening) or expanding the domain of knowledge (widening) [2].

We plan to implement a meta search-engine that will provide users with pages that present these characteristics so that the user is greatly helped in his/her navigational path in the health domain and can focus on the analysis of pages that are of real interest.

REFERENCES

ABOUT THE AUTHORS
Marco Alfano, PhD, Anghelos Centre on Communication Studies and Dipartimento di Matematica e Informatica, University of Palermo, Palermo, Italy, Phone: +39 091 341791, E-mail: marco.alfano@anghelos.org.

Assist. Prof. Biagio Lenzitti, Dipartimento di Matematica ed Informatica, University of Palermo, Palermo, Italy, Phone: +39 091 23891101, E-mail: biagio.lenzitti@unipa.it.

Assist. Prof. Giosuè Lo Bosco, Dipartimento di Matematica ed Informatica, University of Palermo and I.E.ME.ST. Istituto Euro-mediterraneo di Scienza e Tecnologia, Palermo, Italy, Phone: +39 091 23891075, E-mail: giosue.lobosco@unipa.it.

Work partially funded by the PON Smart Cities PON04a2_C “SMART HEALTH – CLUSTER OSDH – SMART FSE-STAYWELL” project.