Relevance of Google-Customized Search Engine vs. CISMeF Quality-Controlled Health Gateway

Jean-François GEHANNO, Gaétan KERDELHUE, Saoussen SAKJI, Philippe MASSARI, Michel JOUBERT, Stéfan J. DARMONI

CISMeF, Rouen University Hospital & LITIS EA 4108, University of Rouen, France
LERTIM EA 3283, University of Marseille, France

Abstract. CISMeF (acronym for Catalog and Index of French Language Health Resources on the Internet) is a quality-controlled health gateway conceived to catalog and index the most important and quality-controlled sources of institutional health information in French. The goal of this study is to compare the relevance of results provided by this gateway from a small set of documents selected and described by human experts to those provided by a search engine from a large set of automatically indexed and ranked resources. The Google-Customized search engine (CSE) was used. The evaluation was made using the 10th first results of 15 queries and two blinded physician evaluators. There was no significant difference between the relevance of information retrieval in CISMeF and Google CSE. In conclusion, automatic indexing does not lead to lower relevance than a manual MeSH indexing and may help to cope with the increasing number of references to be indexed in a controlled health quality gateway.

Keywords. health care quality, access, and evaluation, information storage and retrieval, information, internet, terminology as subject

1. Introduction

The Internet and in particular the Web has become an extensive health information repository. In this context, several quality-controlled health gateways have been developed [1]. Quality-controlled subject gateways were defined by Koch as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery [2]. Among several quality-controlled health gateways, CISMeF ([French] acronym for Catalog and Index of French Language Health Resources on the Internet) was designed to catalog and index the most important and quality-controlled sources of institutional health information in French in order to allow end-users to search them quickly and precisely [1].

From 1995 to 2002, CISMeF content was exclusively manually indexed by a team of four indexers, which are medical librarians and systematically checked by the chief information scientist. Its URLs is http://www.cismef.org. CISMeF uses two standard tools for organizing information: the MeSH thesaurus, including several enhancements,
as previously described [1], and several metadata element sets, in particular the Dublin Core metadata format [3].

Since 2002, automatic indexing tools were developed by the CISMeF team using primarily natural language processing (NLP) and K-nearest neighbors (KNN) methods, followed by a simpler bag of words algorithm [4, 5]. The latter was successfully evaluated in the context of teaching resources [5]. Then the CISMeF team has decided to use this algorithm in the daily practice for most of the Internet resources. Three levels of indexing were defined in the CISMeF catalogue: (a) level 1 or Core-CISMeF (N=18,356): totally manually indexed resources (e.g., guidelines); (b) level 2 or supervised resources (N=5,949): resources that do not need in-depth indexing (e.g., technical reports, teaching resources designed at the national level, document for patients from medical specialties). Supervision means that these resources are primarily indexed automatically, and then this indexing is reviewed by a CISMeF medical librarian; (c) level 3 or automatically indexed resources (N=17,809). The CISMeF editorial board has rated these resources as less important than level 1 and level 2 (e.g., teaching resources designed at the medical school level). Then, the CISMeF wanted to create a level 4, which could be defined as the exhaustive automatically indexed pages from the CISMeF publishers’ websites. The latest can be defined as the websites or publishers that have at least one resource included in the CISMeF catalogue. We used the “Google™ Custom Search Engine” (Google CSE) as Google technology is the most widely used [6].

The goal of this article is to describe how a Google™ custom search engine fitting the CISMeF editorial policy was set-up and to compare the relevance of its results to those of the already existing catalogue.

2. Material and Methods

2.1. CISMeF Information Retrieval Algorithm

The CISMeF terminology is based on the French version of the MeSH thesaurus. To match as much as possible queries with the CISMeF corpus, we have implemented three-step heuristics largely inspired by the PubMed heuristics developed to access the MEDLINE bibliographic database [7]. The process consists in recognizing the user query expression.

Step 1. The reserved terms or the document’s title: If the user query expression matches a term of the CISMeF terminology or the document's title, the process stops, and the answer of the query is the union of the resources that are indexed by the term, and the resources that are indexed by the terms it subsumes, directly or indirectly, in all the hierarchies it belongs to. Stemming, removal of stopwords and a bag-of-words algorithm are used to improve the mapping of the query against the French version of MeSH terms.

Step 2. The CISMeF metadata: The search is performed over all the other fields of the CISMeF metadata set (abstract, author, publisher, identifier ...).

Step 3. Adjacency in the plain texts: A plain text search over the documents with adjacency of \( n \) words with \( n = 10 \times (\text{number of words of the query} – 1) \) is realised.

Similarly to the PubMed/MEDLINE bibliographic database, CISMeF displays the results of a query starting with the most recent document. Contrarily to PubMed/MEDLINE, for the Step 1, the CISMeF heuristics first displays the resources
indexed with the MeSH Major headings, which express the main topic and then the resources indexed with the MeSH minor headings. Therefore, contrarily to what is usually done using search engines, the end-users are not supposed to limit his/her investigation to the first page of results.

2.2. Google-CISMef Customized Search Engine

GoogleTM Custom Search Engine was released in October 2006. It permits to establish which websites to search, prioritize or ignore based on URLs and domain names. In order to fulfill the CISMef editorial policy a whitelist of trustworthy publishers was set up. The custom search was restrained to the URLs and domain names (N=3,952) corresponding to these publishers excluding all others websites. As CISMef is focusing on institutional resources, these publishers are mainly governments from French-speaking countries, national health agencies, medical societies, and universities, especially medical schools. The CISMef editorial board decides the inclusions into the CISMef catalogue. Because a search engine wraps an entire Web site, the corpus of the customized search engine includes the whole corpus of the CISMef health gateway, plus the documents not manually selected by CISMef. Unfortunately the number of pages included in the customized search engine was not available at the time of our evaluation. An approximate number of pages indexed was later included in the Google CSE service. The number was 1.4 million in April 2009.

The CISMef customized version of Google CSE can be searched in two ways: a stand alone approach (http://www.chu-rouen.fr/documed/cismefgoogle.htm) or an integrated approach: when the end user types his query in the CISMef search engine, using a contextual hyperlink he can also send the same query to Google CSE.

2.3. Evaluation

To evaluate the relevance of the information retrieval in CISMef and Google CSE, a list of 50 queries elaborated by physicians from the French Medical Virtual University were used [8]. These 50 queries were elaborated to reflect the types of queries that a medical student usually asks for (e.g., signs and symptoms for asthma). These queries were using free text and not the MeSH controlled-vocabulary used in CISMef. The first parameter was the number of queries without any result for the two systems. The second parameter was a qualitative assessment of the relevance of information retrieval. From the 50 queries, 15 queries were randomly chosen for a more thorough assessment. For each of these 15 queries, the Top 10 answers were evaluated by two physicians from the LITIS Lab (JFG & PM). Even if the binary scale is the paradigm in Information Retrieval research, we have decided to use a 5-point Likert scale (very relevant, relevant, intermediate, irrelevant, and very irrelevant) to rate each resource proposed by both search engines as a result of the query. We decided to use a 5-point Likert scale to obtain accurate results. To limit bias, these two physicians did not belong to the CISMef indexing team, though being members of the same lab. The physicians were blinded regarding the two search engines (CISMef & Google CSE). To compare the two systems, we used the Mann-Whitney test, also named Wilcoxon’s rank sum test, to take into account the ordering of the 5-point Likert scale, and the Wilcoxon’s signed rank test to compare the two evaluators (StatXact-7, Cytel Software Corporation).
3. Results

Google CSE provided at least one page for each of the 50 queries. CISMeF provides at least one resource for only 48 of them. There was no significant difference between CISMeF and Google CSE in terms of relevance of the retrieved information for each of the two evaluators, as displayed in Tables 1 and 2 (Mann-Whitney test; p=0.69 for evaluator A and p=0.10 for evaluator B). For the retrieved document, there was a significant difference between the two evaluators, evaluator B being consistently more severe than evaluator A (p < 0.0001 for Google CSE and p < 0.0001 for CISMeF) (see Table 3). Nevertheless, the two evaluators fully agreed in 42% of their ratings and had less or equal than one point in the Likert scale in 69% of their ratings. Overall, Google CSE and CISMeF proposed respectively 26% to 41% and 30% to 48% of partly or fully irrelevant documents. Among the 15 queries of this study, 12 were recognized as Step 1 in CISMeF (reserved terms or the document’s title), 1 as Step 2 (CISMeF metadata) and 2 as Step 3 (plain text search). Among the results displayed by Google CSE, most of the resources (86%) were not present in the CISMeF catalog. The CISMeF chief medical librarian has analyzed retrospectively if these resources should have been present in the CISMeF catalog. Some resources rated as very relevant by the two evaluators would have not been included in the CISMeF catalog because they were not fulfilling the CISMeF quality criteria.

Table 1. Relevance of CISMeF and Google CSE for evaluator 1

<table>
<thead>
<tr>
<th></th>
<th>V.Rel*</th>
<th>Rel*</th>
<th>Int*</th>
<th>Irr*</th>
<th>V. Irr*</th>
</tr>
</thead>
<tbody>
<tr>
<td>N %</td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
</tr>
<tr>
<td>Google</td>
<td>66</td>
<td>50%</td>
<td>18</td>
<td>14%</td>
<td>14</td>
</tr>
<tr>
<td>CISMeF</td>
<td>65</td>
<td>49%</td>
<td>19</td>
<td>14%</td>
<td>9</td>
</tr>
</tbody>
</table>

*V.Rel: Very relevant; Rel: relevant; Int: intermediate; Irr: irrelevant; V. Irr: very irrelevant

Table 2. Relevance of CISMeF and Google CSE for evaluator 2

<table>
<thead>
<tr>
<th></th>
<th>V.Rel*</th>
<th>Rel*</th>
<th>Int*</th>
<th>Irr*</th>
<th>V. Irr*</th>
</tr>
</thead>
<tbody>
<tr>
<td>N %</td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
</tr>
<tr>
<td>Google CSE</td>
<td>31</td>
<td>23%</td>
<td>22</td>
<td>17%</td>
<td>25</td>
</tr>
<tr>
<td>CISMeF</td>
<td>21</td>
<td>16%</td>
<td>23</td>
<td>17%</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 3. Rating differences between evaluator A and evaluator B (number of Likert point discrepancies among the two evaluators)

<table>
<thead>
<tr>
<th></th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CISMeF</td>
<td>2</td>
<td>12</td>
<td>27</td>
<td>33</td>
<td>47</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Google CSE</td>
<td>3</td>
<td>7</td>
<td>23</td>
<td>27</td>
<td>58</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
4. Discussion

This study focused on a customized version of a search engine vs. a quality-controlled health gateway, being based on the same institutional publishers. The coverage of the customized Google search engine limited to CISMeF publishers is slightly better than the CISMeF catalogue (respective coverage of 100% and 96%). For both evaluators, there was no significant difference between the relevance of the retrieved documents in CISMeF and Google CSE, although there was a tendency in favor of Google CSE for the evaluator 2 (p=0.10). This study has two structural biases against CISMeF: (a) Only the first ten results were evaluated, though no specific algorithm is used to make them the most relevant as the results are ordered mainly on publication date (b) the queries were using free text and did not use the MeSH controlled-vocabulary used in CISMeF, which should have, if correctly used, an added value for end-users.

This study needs further investigation using a higher number of evaluated resources and evaluators in order to confirm or infirm these results. Other search engine technologies could also be accessed. Nevertheless, our results imply that quality-controlled health gateways based on manually indexed resources and vocabulary-controlled thesaurus do not perform better than a totally automated customized search engine when considering only the first results provided.

This does question the use of a controlled vocabulary (such as the MeSH thesaurus), for the relevance in information retrieval and for training purposes. An explanation would be the low consistency of indexing between experts. A very preliminary result based on 10 resources show a 30% consistency. We still expect manual indexing to be better than Google-CSE search, when the nature of the document is very important: For queries mixing content and the nature of the document e.g., guidelines for asthma, we assume than a quality-controlled health gateway should outperform a customized search engine. We plan to elaborate the next evaluation of information retrieval on mixed queries. Combining expertise of both the CISMeF quality-controlled health gateway and a customized version of generic search engine integration may increase the satisfaction of health professionals and citizens when they look for reliable health information on the Web.

References