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

Jean-François Gehanno\textsuperscript{a}, Gaétan Kerdelhué\textsuperscript{a}, Saoussen Sakji\textsuperscript{a}, Philippe Massari\textsuperscript{a}, Michel Joubert\textsuperscript{b}, Stéfan J. Darmoni\textsuperscript{a}

\textsuperscript{a} CISMeF & TIBS, LITIS Lab
Rouen University Hospital & Rouen Medical School, France

\textsuperscript{b} LERTIM EA 3283, University of Marseille. France

Email: Stefan.Darmoni@chu-rouen.fr

MIE August 2009
Introduction

- Quality-controlled subject gateways were defined by Koch as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery.

- 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:
  - began in *February 1995*
  - [www.cismef.org](http://www.cismef.org)
  - N= 12: 3.5 librarians, 1.5 medical informaticians, 1 computer scientist (junior lecturer), 3 engineers, 3 PhDs
CISMeF terminology

- Two standard tools for organising information:
  - the MeSH (Medical Subject Headings) thesaurus from the US National Library of Medicine
  - Several metadata element sets
    - the Dublin Core metadata format + CISMeF specific fields
    - For teaching resources, IEEE 1484 LOM metadata format
      11 elements of the LOM Educational category => DC.Education
    - For evidence-based medicine resources, CISMeF specific fields: level of evidence + method to evaluate it

DC-2004, International Conference on Dublin Core and Metadata Applications

CISMeF Information Retrieval

- Since 2005, three levels of indexing in CISMeF
  - Level 1: manuel indexing (e.g. guidelines) \( (N=18,356) \)
  - Level 2: supervised indexing (e.g. technical report or teaching document from national medical societies) \( (N=5,949) \)
  - Level 3: automatic indexing (e.g. SCPs, teaching document from one medical school) \( (N=17,809) \)

- Wish of level 4
  - exhaustive automatically indexed pages from the CISMeF publishers
  - Instead of reinventing the wheel
    - "Google™ Custom Search Engine" (Google CSE), using the "Google Co-op™ platform"
Objective

- To describe and to evaluate the cooperation between
  - the CISMeF quality-controlled health gateway and
  - a customized version of a generic search engine from Google
    - "Google™ Custom Search Engine" (Google CSE), using the "Google Co-op™ platform
Methods: current IR in CISMeF

- Only three steps
  Step 1: Reserved terms ($\in$ CISMeF terminology) OR document's title

  Step 2: *The CISMeF metadata*
  Mixing the reserved terms, all fields and adjacency in the titles (word adjacency: $(n-1)*5$)

  Step 3: *Adjacency in the plain texts*
  Mixing the reserved terms, all fields and adjacency in the plain texts (word adjacency: $(n-1)*10$)

Methods: Google-CISMeF CSE

- Possible to define a customized version of Google on the basis of the common Google crawler
- Providing a list of trustworthy web sites from the CISMeF database (N=3,952) => 1M pages
- These publishers are mainly
  - governments from French-speaking countries
  - national health agencies (e.g. Haute Autorite de Sante in France),
  - medical societies, and
  - universities, especially medical schools
Methods: Google-CISMeF CSE

- Google CSE allows adding generic health metadata (e.g. guidelines)
  - at the **publisher** level and
  - not at the **resource** level as it is done in the CISMeF catalogue.

- It is also possible to add specific health metadata:
  - in this work, three metadata based on the **target** of the Web site:
    (a) health professional,
    (b) students and
    (c) patients and lay people.

- Google CSE displays the results of a query, using the Google Page Rank Algorithm,

- The CISMeF customized version of Google CSE can be searched in two ways:
  - a stand alone approach (URL: [http://www.chu-rouen.fr/documed/cismefgoogle.htm](http://www.chu-rouen.fr/documed/cismefgoogle.htm)) or
  - an integrated approach (knowledge coupling) from CISMeF search engine and terminology browser
Evaluation

- To evaluate the relevance of the information retrieval in CISMeF and Google
  - 50 queries elaborated by physicians from the French Medical Virtual University were used
- These queries were using free text and not the MeSH controlled-vocabulary used in CISMeF.
- First parameter = number of queries without any result for the two systems
- Second parameter = qualitative assessment of the relevance of information retrieval
  - 15 queries out 50 were randomly
  - Top 10 answers evaluated by two physicians from the LITIS Lab (JFG & PM).
Evaluation

- Assessment using a 5-point Likert scale (very relevant, relevant, intermediate, irrelevant, and very irrelevant)
- To avoid bias, these two physicians did not belong to the CISMeF indexing team
- The physicians blinded regarding the two search engines (CISMeF & Google CSE)
- Mann-Whitney test, also named Wilcoxon's rank sum test, and the Wilcoxon's signed rank test to compare the two evaluators
- Manually evaluated the precision of the Top 20 answers of queries 4 & 5
- Consensus of two authors
Results

- Coverage
  - Google CSE provided at least one page for each of the 50 queries; CISMeF N=48

- Relevance
  - No significant difference between CISMeF and Google CSE in terms of relevance of the retrieved information for each of the two evaluators (Mann-Whitney test; p= 0.69 for evaluator A and p=0.10 for evaluator B)
  - Significant difference between the two evaluators, evaluator B being consistently more severe than evaluator A (Wilcoxon's signed rank test: p < 0.0001 for Google CSE and p < 0.0001 for CISMeF)
  - 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
  - Among the results displayed by Google CSE, most of the resources (86%) were not present in the CISMeF catalog
  - 15 queries of this study, 12 were recognized as Step 1 in CISMeF, 1 as Step 2 and 2 as Step 3
## Results

### 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></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Google CSE</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>

### 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></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</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>
Discussion

- Slightly better coverage for Google CSE vs. CISMeF (100% vs. 96%)
- No significant difference between the relevance of the retrieved documents in CISMeF and Google CSE
  - tendency in favor of Google CSE for the evaluator 2 (p=0.10)
  - surprising for the CISMeF team, and especially for the four medical indexers
    - expecting a significant better relevance of retrieved documents for CISMeF, which is partially manually indexed vs. Google-CSE, which is totally automatically indexed
Discussion

- This study has three structural biases against CISMeF:
  - (a) in CISMeF, the first 10 documents were displayed according to their date of publication as it is currently the case in PubMed.
  - (b) we made the hypothesis that most of the end-users are using CISMeF as a search engine and do not go beyond the first page.
  - (c) the queries were using free text and did not use the MeSH controlled-vocabulary used in CISMeF.
  - (d) performance of Google CSE could be partly due to its greater collection size ($10^6$ vs. $10^5$).
Since 2009, four levels of indexing in CISMeF

- Level 1: manuel indexing (e.g. guidelines)
- Level 2: supervised indexing (e.g. technical report or teaching document from national medical societies)
- Level 3: automatic indexing (e.g. SCPs, teaching document from one medical school)
- Level 4: extending the CISMeF corpus => Google CISMeF (restricted to publishers included in CISMeF)
Changes in CISMeF information retrieval

- Since 2009, CISMeF is fully « multi-terminological »
  - CISMeF backoffice contains the main health terminologies available in French (e.g. SNOMED Int, ICD10, ATC, CCAM)
  - Multi-terminological automatic indexing (better recall)
  - Multi-terminological information retrieval

- Modification of the IR ranking algorithm
  - MeSH Major (or Title) first (display of score)
    - Then, date (as PubMed)
  - Automatic (Title or SubTitle)
  - Minor MeSH