What about trust in a question answering system?

S. Cruchet, A. Gaudinat, C. Boyer, T. Rindflesh
Sarah.Cruchet@healthonnet.org

Why a Question / Answering system?

Keywords system

What is diabetes?

1,430,000 documents

Which document contains my answer?

documents ≠ answers

Why in the health domain?

Many health documents on the Internet. The answer exists but no system proposes direct answers to a given query.

The health domain is very important for the citizen. 1 in 5 patients use the Internet to search for health information.

Hypothesis

We assume that the HONcode documents are more trustworthy than noncertified documents because such certified Websites abide by the 8 HONcode principles of ethics and quality.

This hypothesis has been assessed by independent researchers for ten common orthopaedic sports medicine diagnoses.

One of our hypotheses is that the reliability of an answer is strongly related to the trustworthiness of the document it belongs to.

The more frequently a statement is repeated, the more it is to be the possible answer.

Materials and Methodology

Study based on the methodology developed for a study conducted at the U.S. National Library of Medicine (NLM) on information retrieval systems.

Comparison of the results of the HON-developed QA system when applied to two different databases:

- QAISON_honcode uses only the database of 7,300 HONcode certified health and medical websites.
- QAISON_google uses only Google results through our QA system.

100 questions in English related to diabetes have been randomly selected from the 136 initial ones used to develop the question analyzer. Samples of questions used:

- What does blood sugar level mean?
- What do I need to know about high blood sugar?
- What is the treatment for diabetes?

Questions are submitted to each database.

Answers are collected and anonymised so it is not possible to know the origin of the results—neither the database nor the URL itself.

Answers are graded by a health professional according to the following scale:

- A+ (very relevant and reliable),
- A (relevant and reliable),
- A (not the whole answer),
- B+ (leading to a response),
- B (may lead to the answer),
- C (not relevant).

We consider an answer relevant if it gets an A or a B.

Results

<table>
<thead>
<tr>
<th>System</th>
<th>QAISON_honcode</th>
<th>QAISON_google</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.59</td>
<td>0.36</td>
</tr>
<tr>
<td>Bpref</td>
<td>0.50</td>
<td>0.34</td>
</tr>
<tr>
<td>R-prec</td>
<td>0.59</td>
<td>0.38</td>
</tr>
<tr>
<td>MRR</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>P@5</td>
<td>0.54</td>
<td>0.36</td>
</tr>
<tr>
<td>P@10</td>
<td>0.32</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Results are globally better for QAISON_honcode.

59% of answers coming from the HONcode database are relevant against 36% for Google.

Out of the first 5 answers found by QAISON_honcode, more than half correspond exactly to the expected answers of the question asked.

QAISON_honcode has 10% less of the first good result than Google (76% compared to 86% for the QAISON_google).

Importance of the first result: it is those which will be read by the user.

Important point:

The Google results include non-HONcode-certified websites as well as HONcode-certified websites.

Discussion

The questions are specific enough to be clearly related to the health domain.

Output performance for the first three types of measures (MAP, Bpref, and R-prec) for information retrieval systems, these measures are more representative of the behaviour of the end user selecting the good results from many others.

Conclusion and Perspectives

Correlation between the reliability of the answers of a QA system with the database of webpages used to retrieve the answers.

In the interface, the user will have access to the answer in its context for better traceability of the response.

The date of the information in the health domain is crucial. So further development will favor the update date of the information.

User interface (www.healthonnet.org/QA)