Using Local Precision to Compare Search Engines in Consumer Health Information Retrieval

Carla Teixeira Lopes  
Departamento de Engenharia Informática  
Faculdade de Engenharia da Universidade do Porto  
Rua Dr. Roberto Frias s/n  
4200-465 Porto, Portugal  
ctl@fe.up.pt

Cristina Ribeiro  
Departamento de Engenharia Informática  
Faculdade de Engenharia da Universidade do Porto/INESC-Porto  
Rua Dr. Roberto Frias s/n  
4200-465 Porto, Portugal  
mcr@fe.up.pt

1. INTRODUCTION

Patients, their family and friends are increasingly using the Web to search for health information [2]. The last Pew Internet report on health information [4] reveals that 61% of the American adults look online for health information. In the Internet users, this proportion rises to 83%. A previous study reported that 66% of health information sessions start at health-specific websites [3]. Large companies in information retrieval have been developing efforts in the health area (e.g. Google Health and Bing Health) and several health-specific services are also appearing.

This study evaluates the performance of 4 generalist SE (Google, Bing, Yahoo! and Sapo) and 3 specific SE (MedlinePlus, WebMD and Sapo Saúde) in health information retrieval. The evaluation is based on the data collected in a user study with undergraduate students and work tasks defined according to the framework proposed by Borlund [1]. Besides an overall comparison, the SE are also compared according to the type of clinical question and medical specialty.

ABSTRACT

We have conducted a user study to evaluate several generalist and health-specific search engines on health information retrieval. Users evaluated the relevance of the top 30 documents of 4 search engines in two different health information needs. We introduce the concepts of local and global precision and analyze how they affect the evaluation. Results show that Google surpasses the precision of all other engines, including the health-specific ones, and that precision differs with the type of clinical question and its medical specialty.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; J.3 [Computer Applications]: Life and Medical Sciences

General Terms: Experimentation, Performance, Human Factors.

Keywords: Health information retrieval, evaluation, precision, user study.

2. METHOD

We conducted a user study with 5 work tasks that were defined based on popular questions submitted to web health support groups. Each work task acts as the context of 4 information needs (IN) that are linked to it. The defined work tasks are available at http://www.carlalopes.com/research/userstudy2.html and are associated with the following medical specialties: gynecology, dermatology, psychiatry and urology. Moreover, each IN is associated with one of the following types of clinical questions: overview, diagnosis/symptoms, treatment, prevention/screening, disease management, prognosis/outcome.

Each user chose 2 IN that, after being transformed into queries, were submitted to 4 SE selected by the user. Following the pooling approach, each user assessed the relevance of the top 30 documents returned by each SE. Totally there were 82 sets of judged documents, one for each pair of user and IN, from which were excluded duplicates. Forty-one undergraduate users participated in this study (27 females; 14 males) with a mean age of 27.2 years (SD=9.8). These users evaluated 9,572 documents, less than 41×2×4×30 because some queries returned less than 30 documents.

To evaluate and compare the SE we used the 11-point interpolated average precision and the Mean Average Precision (MAP). In these measures we used two types of precision, a local precision that is calculated based on the actual relevance judgments of the users and a global precision, based on the set of relevant documents assembled for the IN in the pool. As human relevance judgments are idiosyncratic, these two types of precision will allow us to compare the real precision - as judged by the users - and the estimated precision - calculated in a test-collection fashion way.

The first type of precision is calculated using a local set of relevant documents that is defined by \(LRel(u, in) = \{doc : doc \in P(u, in) \land RJ(doc) = 1\}\). The second uses a global set of relevant documents that is defined by \(GRel(in) = unique(\bigcup_{u \in U} LRel(u, in))\). In these formulas, \(u\) is an user; \(U\) represents the set of all users, \(in\) is an information need; \(doc\) is a document; \(P(u, in)\) is the pool of judged documents for user \(u\) and information need \(in\) and \(RJ(doc)\) is the relevance judgment for document \(doc\) that can be 0 or 1.

3. RESULTS

With the exception of one IN, all information needs were associated with at least one user. All users chose Google as one of the four SE. The other SE with more selections were
the Sapo Saúde (27 users), Bing (25 users) and MedlinePlus (23 users).

Table 1 presents the Local MAP, calculated with local precision, the Global MAP, calculated with global precision, and the maximum recall of each SE. As global map is more comprehensive, it was expected to have a Local MAP lower than Global MAP, and the difference is in fact significative (p<0.01). From this table we can see that Google is the engine with larger MAP and maximum recall. Although the literature mentions that the variation of individual assessors’ judgements have little impact on the the relative effectiveness ranking of different systems [5], we found that the use of these two types of precision slightly changes this ranking. MedlinePlus and Sapo Saúde, two health-specific SE, appear in both rankings but in reverse order, after Google. Yahoo! is the SE that rises more positions, from 6th to 3rd, in the ranking with local MAP.

Recall is not a uniform measure because each SE has its own collection. Moreover, while generalist SE index contents from several sources, the 3 analyzed health-specific SE only index their own contents. Therefore, the maximum recall in GRel(in), presented in Table 1, is not supposed to be used as a measure to rank the 7 SE. In the generalist SE, we can see that in terms of recall, Bing, Sapo and Yahoo, in this specific order, follow Google. Considering the unfair situation of health-specific SE, an individual analysis of their recall make us predict that, in a pool composed by only their documents, they would have a good recall. The recall issues described above and space constraints made us decide to omit the 11-point interpolated average precision analysis.

We found significative differences in local and global MAP between SE, both with p<0.01. As local MAP is more strict on users’ relevance judgements, it is more realistic and therefore will be used in the following analysis. We further studied the differences between each pair of SE and found that Google has a significant larger MAP than the other SE (p<0.05 with Sapo and p<0.01 in the other comparisons). The differences between other SE are not significative.

Our analysis has also focused on the type of clinical question and on the medical specialty of the query. In both cases, we have compared the MAP of each category of the query type and query specialty, the MAP of each SE in each category and the MAP of each category in all SE.

Each IN is associated with one type of clinical question. The small number of submitted queries (<5) of the Disease Management and Prognosis/Outcome categories, made us omit them from our analysis. In the overall analysis, we found significant differences at α=0.1 (p=0.06) between types of queries. The Overview query type has the largest MAP median and the smallest dispersion. A similar analysis at an engine perspective, show us that only MedlinePlus and Yahoo have significant differences in MAP at query types at α=0.1 (p=0.08 and p=0.07, respectively). In MedlinePlus, the Overview and Treatment query types have significative larger MAP at different α (p=0.01 and p=0.07, respectively) than the Prevention/Screening ones. In Yahoo, the Overview category is significantly better in MAP than Diagnosis/Symptoms and Treatment (p=0.07 and p=0.05, respectively). In the comparative analysis of the engine’s precision in each query type, in the Prevention/Screening query type, the engines have significant different MAP (p=0.06). In a deeper analysis, we found that in this type of question, Google is significantly better (p<0.05) than MedlinePlus, Sapo and WebMD and also that WebMD is significantly worse than Bing (p=0.03) and MedlinePlus (p=0.03).

In the overall analysis on the medical specialty, we found significant differences at α=0.1 (p=0.06). At different significance levels, dermatology has lower MAP than gynecology (p=0.01), psychiatry (p=0.00) and urology (p=0.05). In an engine analysis, only Sapo Saude has significative differences between specialties (p=0.09). A deeper analysis show us that in this engine, the psychiatry questions have higher MAP than urology (p=0.09), gynecology (p=0.05) and dermatology (p=0.01). In an specialty analysis, we could verify that there are significant differences between engines in the psychiatry (p=0.04) and dermatology (p=0.08) specialties. In psychiatry, we found that Google has a significant better MAP than Bing (p=0.00), MedlinePlus (p<0.05), Sapo (p=0.04) and Yahoo (p=0.02). On the other hand, Bing has also a worse MAP than MedlinePlus (p=0.02), Sapo Saude (p=0.00) and WebMD (p<0.04). In dermatology, Google has a significant better MAP than Bing (p=0.05), MedlinePlus (p=0.05) and Sapo Saude (p=0.05).

4. CONCLUSION

We have introduced two types of precision, a local and a global one. The former is closer to each user relevance judgments and the latter is similar to the notion of precision used in TREC evaluations. We found that there are significative differences between them and that different ranking of search engines appear with these two type of precisions. Future work will be done on the influence of context features in relevance evaluations by human assessors and on ways to deal with different judgments. We also found that generalist search engines don’t have lower precision than health-specific ones. In fact, Google has shown a precision significantly higher than all the others search engines. It would be interesting to complement this study with an evaluation of the documents’ contents by health experts and to analyze its correlation with user’s judgements.

5. REFERENCES