Comparative evaluation of web search engines in health information retrieval

Carla Teixeira Lopes, DEI, Faculdade de Engenharia, Universidade do Porto, ctl@fe.up.pt
Cristina Ribeiro, INESC Porto, DEI, Faculdade de Engenharia, Universidade do Porto, mcr@fe.up.pt

This is a post-print of an article accepted for publication in the *Evaluating Web Search Engines* Special Issue of Online Information Review (Emerald) @ 2011.

Purpose: With this work we intend to evaluate several generalist and health-specific search engines on retrieval of health information by consumers. We compare the retrieval effectiveness of these engines in different types of clinical queries, medical specialties and condition's severity. Finally, we compare the use of evaluation metrics for binary relevance scales and for graded ones.

Design/methodology/approach: We conducted a user study in which users evaluated the relevance of documents retrieved by 4 search engines in 2 different health information needs. Users could choose between generalist (Bing, Google, Sapo and Yahoo!) and health-specific search engines (MedlinePlus, SapoSaúde and WebMD). We then analyse the differences between search engines and groups of information needs with six different measures: graded average precision (gap), average precision (ap), gap@5, gap@10, ap@5 and ap@10.

Findings: Results show that generalist web search engines surpass the precision of health-specific engines. Google has the better performance, mainly on the top-10 results. We found that information needs associated with severe conditions are associated with higher precision just like overview and psychiatry questions.

Originality/value: Our study is one of the first studies to use a recently proposed measure to evaluate the effectiveness of retrieval systems with graded relevance scales. It includes tasks of several medical specialties, types of clinical questions and different levels of severity what, to the best of our knowledge, has not been done before. Also, it is a study in which users have a large involvement in the experiment. Results are useful to understand how search engines differ in their responses to health information needs, to inform about what types of online health information are more common on the Web and to infer ways to improve this type of search.

Keywords: Evaluation, Health information retrieval, User study, Graded-relevance, Web search engines

Paper type: Research paper

1. INTRODUCTION

Patients, their family and friends, commonly designated by health consumers, are increasingly using the Web to search for health information. The last Pew Internet report on health information (Fox and Jones, 2009) 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 generalist search engines and 27% start at health-specific websites (Fox, 2006). 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 have been appearing.

According to Hersh (2008), the amount and quality of evaluation research didn't follow the changes that Information Retrieval has suffered with the ubiquity of the Web. In his opinion, the number of studies that evaluate the performance of web search systems in health is "surprisingly small". We focus on consumers because, on health information retrieval, they receive less attention when compared to professionals (Lopes and Ribeiro, 2010).

This study evaluates the performance of 4 generalist search engines (Google, Bing, Yahoo! and Sapo) and 3 health-specific search engines (MedlinePlus, WebMD and SapoSaúde). 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 (Borlund, 2003). Besides an overall comparison, search engines are also differentiated by their performance on different clinical questions, medical specialties and levels of severity.

We start to review previous works on evaluation of web search engines and, more specifically, on their evaluation on the health domain. Next, we describe our methodology, present the study and discuss our results. We conclude with some final remarks.

2. LITERATURE REVIEW

Evaluation in Information Retrieval

Information Retrieval is a highly empirical field in which evaluation is essential to demonstrate the performance of new techniques (Manning et al., 2008). The use of test collections is the dominant evaluation standard, being used since the early 1950s along with evaluation measures (Sanderson, 2010). Since 1992, TREC (Text REtrieval Conference) has been a major forum to discuss research evaluated through this model. The use of test collections is particularly well suited to system-oriented performance evaluations that focus on specific aspects of system.

Experimental methods involving the user also exist and have been promoted by Ingwersen and Järvelin (2005) and by Borlund (2003). Ingwersen (2009) identifies three major types of research methods involving users: ultra-light IR interaction experiments, interactive light IR experiments and naturalistic IR field studies in the context of, for instance, an organizational setting. The first focus on short-term IR interaction composed of 1 to 2 retrieval runs. The second entails session-based multi-run interaction with more intensive monitoring like log analysis, interviews and observation. These studies can be run in a laboratory, in naturalistic settings or in the Internet through what Sanderson (2010) calls Live Labs. Another method that has been growing since the appearance of web search engines involves the study of user behaviour using query logs.

The two most popular measures for IR effectiveness are precision and recall (Manning et al., 2008). Precision is the fraction of retrieved documents that are relevant and Recall is the fraction of relevant documents that are retrieved. Along with the F measure that is the weighted harmonic mean of precision and recall, these are the most used measures in unranked retrieval. In a ranked retrieval context, precision-recall curves can be plotted. A very common measure is the Mean Average Precision (MAP) and, in scenarios like the Web in which it is important to have good results on the first pages, precision is also measured at fixed levels of retrieval (e.g.: precision at 10). In situations where non-binary scales of relevance are used, Normalized Discounted Cumulative Gain (NDCG) is popular. This measure considers that highly relevant documents are more valuable than marginally relevant ones and their value decreases as the position in the ranking increases (Järvelin and Kekäläinen, 2002). Very recently, Robertson et al. (2010) proposed a new measure named Graded Average Precision (GAP) that generalizes average precision to multi-graded relevance.

Evaluation of Web Search Engines

Although this paper focuses on the health domain, we present a brief overview of previous works that aim to evaluate and compare the performance of generalist web search engines. In Table 1 we present a list of research papers with these goals, along with the number of search engines evaluated and the type of measures used to compare them. All studies, except the one from Shang and Li (2002), involved users in their evaluation, either to define the information needs, the queries or to judge the relevance of the documents. Shang and Li (2002) computed relevance scores using three traditional algorithms (cover density ranking, Okapi similarity measurement and vector space

model) and an additional one developed by the authors. Besides the differences presented in Table 1, other distinctions lay in the users characteristics, the information needs, the queries generated (e.g.: number, how, any restrictions?) and the method used to judge results (e.g.: number of results judged, by whom, relevance scale). Contrasting the other studies, Vaughan's work (2004) uses a continuous relevance scale (from the most relevant to the least relevant result) instead of a discrete relevance scale. Statistical comparison of the measures was the standard method to analyse the results.

Table 1 – Previous studies on Evaluation of Web Search Engines

Study	#Search Engines	Evaluation criteria
(Chu and Rosenthal, 1996)	3	Search capabilities, output options, documentation and interface. Response time, precision.
(Gordon and Pathak, 1999)	8	Recall and precision at varying numbers of retrieved documents.
(Gwizdka and Chignell, 1999)	3	Precision, presentation, user effort and coverage.
(Hawking et al., 2001)	20	Precision oriented measures: P@n at n≤20, mean reciprocal rank of first relevant document and TREC-style average precision.
(Shang and Li, 2002)	6	Precision oriented measures.
(Su, 2003b)	4	16 performance measures in 5 criteria: relevance, efficiency, utility, user satisfaction and connectivity.
(Tang and Sun, 2003)	4	First 20 full precision, search length and rank correlation.
(Vaughan, 2004)	3	Quality of result ranking, ability to retrieve top ranked pages and 3 stability measurements.
(Lewandowski, 2008)	5	Precision measures and recall-precision graphs applied to results and to their descriptions.

The work from Chu and Rosenthal (1996) also includes an evaluation methodology for web search engines. In their opinion, search engines should be evaluated considering 5 aspects: composition of web indexes, search capability, retrieval performance, output option and user effort. Su, in a work previous to the one presented in Table 1 proposes a set of criteria and measures as a systematic model to evaluate web search engines (Su, 2003a).

Evaluation of Web Search Engines in the health domain

Hersh (2008) does a broad review of studies that evaluate search systems in the health domain in terms of system and user performance. The majority of the studies focus on professionals' systems, mainly using MEDLINE. The number of web searching systems evaluated in the literature is, as Hersh says, "surprisingly small". In this section we will review previous studies that, at least, evaluate one web search engine. Studies that, for example, evaluate and compare two MEDLINE search systems will not be reported here. We will give more attention to papers that focus on health consumers.

That we are aware of, only three papers explore the performance of web search engines in the health domain when used by professionals. This might be explained by these users' preference on sources like MEDLINE instead of the Web to satisfy their

information needs. In Table 2 we present the web search engines included in each study and also their evaluation criteria. All studies compare web search engines with other kind of resources. In the study from Johnson et al. (2008), users were randomly assigned to Google or other web resource of their choice. Graber et al. (1999) selected 10 questions posed by physicians and Yu and Kaufman (2007) choose 12 physicians questions in the format "What is X?". On the other hand, the other work (Johnson et al., 2008) used 10 medical questions extracted from a multiple choice exam. All papers evaluated the medical quality of the contents retrieved and the number of links used to get to the answer. A few other criteria were used, as can be seen in Table 2. Results of these studies report that health specific search engines behave poorly when compared with generalist engines. In studies where Google was used, authors concluded that this is an effective engine for medical information.

Table 2 – Studies evaluating Web Search Engines in the professional health domain

Study	Search Engines	Evaluation criteria
(Graber et al., 1999)	1 site,4 generalist engines,9 medicine-specific engines,2 medical meta-lists	Number of questions answered, correctness of the answers, number of links followed to get an answer and how well documented the answer was using Health on the Net criteria.
(Yu and Kaufman, 2007)	Google,MedQA, Onelook,PubMed	Quality of answer, ease of use, time spent, and number of actions taken.
(Johnson et al., 2008)	Google vs. other web resources	Resource efficiency (inversely related to number of links used to identify the correct answer) and correctness (#correct answers/#answered questions).

We have analysed seven papers that evaluate web search engines on the health domain in the consumer's perspective. A summary of the main differences between these works is presented on Table 3.

All papers evaluate and compare several search engines and, if we exclude the work from Jones and Tim (2008), all include, at least, one generalist search engine. The works from Kumar (2005) and Tang et al. (2006) are user studies. Only the Wu and Li (1999)'s study had the contribution of two librarians. As seen in Table 3, the authors, either selecting questions posed to librarians/clinicians or consumers' popular questions, formulated almost all information needs. The method used to evaluate popularity was not mentioned in any of the papers. Bin and Lun's (2001) work considers two search types: single keyword searches (SKS) in which the authors want to retrieve information about a term and question-answering (QA) to evaluate the answer to a clinical question. In QA the authors used the questions defined in the work of Graber et al. (1999). Tang et al. (2006) employed two types of information needs, related and non-related to treatments. The first type of queries is based on treatments' names to which they had evidence ratings. This allowed the evaluation of the quality of the health information retrieved without health professionals. The second type of queries was extracted from search logs of a depression search engine and from the suggestions given by a tool based on common queries. In the first type of queries, user judgments were made on documents' relevance and on the treatment recommendation in retrieved documents. In the work of Kumar (2005), the relevance of the documents was assessed by users in an aggregated way in a post-search questionnaire. Health professionals judged document content quality in an individual way.

Table 3 – Studies evaluating Web Search Engines in the consumer health domain

		(Wu and Li, 1999)	(Bin and Lun, 2001)	(Ilic et al., 2003)	(Kumar, 2005)	(Tang et al., 2006)	(Jones and Timm, 2008)	(Knight et al., 2009)
Search Engines	#	7	8	9	3	4	6	9
	Which? (generalist signed with *)	Altavista* Excite* Hotbot* Infoseek* Medical World* Northern Light* Yahoo!*	Altavista/Health Excite/Health HardinMD MedHunt MediAgent Medical World Medical Matrix Yahoo!/Health	AltaVista* DrKoop Excite* HealthInsite HON Google* MedlinePlus NHS Yahoo!*	Google* Healia MedHunt	4sites Blue Pages (BPS) Google* HealthFinder	Healia Healthfinder Healthline MedlinePlus Medstory Yahoo!/Health	Dogpile* Healia Healthline Google* Jux2* Kosmix Health RevolutionHealth WebMD Yahoo!*
User	#	2	-	-	66	Not specified	-	-
	Type	Librarians	-	-	Volunteers	Research assistants	-	-
Inf. Needs	#	5	SKS: 4, QA: 8	1	6	Not specified	Not specified	5
	How?	Questions posed to librarians	SKS:not specified QA:questions posed to clinicians	Not specified	Popularity	Not specified	Not specified	Popularity
	By whom	Authors	SKS:authors; QA: (Graber et al., 1999)	Authors	Authors	Not specified	Not specified	Authors
	Categ.	5	-	-	-	2:treatment (T) and others (O)	-	-
Queries	#	5	SKS:4; QA:8	20	Not specified	101	Not specified	5
	How?	Keywords linked by operators	Not specified	Phrases and Boolean	Not specified	T:treatments' names;O:BPS logs +suggestion tool	Not specified	Following the characteristics of popular queries
	By whom?	Librarians	Authors	Authors	Users	Authors	Not specified	Authors
	#judgm.	30	SKS:100;QA:5	50	Not specified	10	Not specified	10
Judgm.	**By whom?**	Librarians	Authors	Authors	Health professionals	Users	Not specified	Authors
	#Levels	2	Not specified	Not specified	10	4	Not specified	Not specified
	#total.	150	440	4927	Not specified	Not specified	Not specified	Not specified

As can be seen in Table 4, the criteria to evaluate the search engines included the relevance of the results, quality of results from a medical point of view, usability and search engines' features. Results' quality is evaluated in different ways. Some analyse documents' characteristics like authorship, evidence of citation, disclosure and currency ((Wu and Li, 1999), (Ilic et al., 2003)). Kumar (2005) asked health professionals to judge the accuracy and trustworthiness of results and Tang et al. (2006) used treatments' evidence ratings to validate their quality. Finally, Knight et al. (2009) used the FA4CT algorithm (Eysenbach and Thomson, 2007) for the same purpose.

Table 4 – Criteria used to evaluate search engines in the consumer health domain

Study	Evaluation criteria
(Wu and Li, 1999)	Relevance (relevant hits per queries topic), source reliability (authorship, source, disclosure, currency), duplicate and inactive links, search engine features.
(Bin and Lun, 2001)	SKS: number of medical resources about the topic. QA: number of questions answered, number of links followed.
(Ilic et al., 2003)	Relevance. Quality (target audience, authorship and evidence citation).
(Kumar, 2005)	Usefulness, ease-of-use, relevance, overall satisfaction, accuracy, trustworthiness.
(Tang et al., 2006)	Relevance (MAP, NDCG). Quality of advice according to EBM (Quality Score).
(Jones and Timm, 2008)	Major Features, Navigation, Timeliness and Quality of Retrieved Items, Search Interface and Strategy, Search Results/Display, Deficiencies or Disadvantages, Overall Effectiveness.
(Knight et al., 2009)	Popularity, usability (SELP & SERP), relevance (precision and relative recall), results quality and features.

A common pattern emerges from all studies that include generalist search engines. All conclude that the performance of generalist search engines is equal (Bin and Lun, 2001), or better (all the others) than the performance of health-specific ones. Regarding information quality, some studies concluded that health-specific ones outperformed the generic ones ((Kumar, 2005), (Tang et al., 2006)) and the other said there were no differences (Ilic et al., 2003).

From all these studies, the one from Jones and Timm (2008) stands out for its more qualitative nature. The work of Tang et al. (2006) is the closest to our study. It is a user study, it has objective methods and it uses well-known measures in Information Retrieval. Our study's main differences stand in the larger involvement of users in the experiment, the use of different measures to evaluate performance and the inclusion of tasks of different medical specialties, types of clinical questions and levels of severity. More specific differences are detailed in the sequel.

3. METHODOLOGY

We conducted a user study with undergraduate students that allowed us to evaluate and compare the performance of generalist and health-specific search engines. Study details are described next.

Search Engines

We included 4 generalist web search engines and 3 health-specific ones in our study, as expressed in Table 5. Google, Bing and Yahoo! were selected for their popularity. At

least in two rankings (alexa.com and hitwise.com) they are positioned as the top-3 search engines. Sapo was included because it is the main Portuguese search engine. MedlinePlus is a service of the U.S. National Library of Medicine and was included for its credibility. We also included WebMD because, according to the US market share of visits (http://www.marketingcharts.com/interactive/top-10-health-medical-information-websites-july-2010-13919/), it is one of the main services of this type and SapoSaúde for the same reason in what concerns Portugal. Sapo and SapoSaúde are both owned by the same company.

Table 5 – Search engines included in this study

	URL	Type
Bing	http://www.bing.com/	Generalist
Google	http://www.google.com/ http://www.google.pt/	Generalist
MedlinePlus	http://www.nlm.nih.gov/medlineplus/	Health-specific
Sapo	http://pesquisa.sapo.pt/	Generalist
SapoSaúde	http://saude.sapo.pt/	Health-specific
WebMD	http://www.webmd.com/	Health-specific
Yahoo!	http://www.yahoo.com/	Generalist

Work tasks

Following the framework proposed by Borlund (2003), we defined five work tasks based on popular (most viewed) questions submitted to web health support groups. Each work task acts as the context of four information needs (IN) that are linked to it. The defined work tasks are available at [omitted for anonymity reasons] and are associated with the following medical specialties: gynaecology, dermatology, psychiatry and urology. They are also categorized as severe or non-severe. Any life threatening or long-term, chronic illness is considered a severe condition.

Each information need is associated with one of the following types of clinical questions: overview, diagnosis/symptoms, treatment, prevention/screening, disease management and prognosis/outcome. These categories were defined upon the categories of clinical questions presented by Hersh (2008) and the information categories available in MedlinePlus topics.

Table 6 - Work tasks used in this study

Task	Specialty	Severe?	IN	Clinical question	#Users[#F,#M]
1	Psychiatry	Yes	IN1.1	Overview	8[5,3]
			IN1.2	Overview	2[0,2]
			IN1.3	Disease Management	2[1,1]
			IN1.4	Treatment	5[4,1]
2	Dermatology	No	IN2.1	Prevention/Screening	8[3,5]
			IN2.2	Prevention/Screening	2[1,1]
			IN2.3	Prevention/Screening	5[3,2]
			IN2.4	Prevention/Screening	0[0,0]
3	Gynaecology	Yes	IN3.1	Diagnosis/Symptoms	7[5,2]
			IN3.2	Diagnosis/Symptoms	6[5,1]
			IN3.3	Treatment	1[1,0]
			IN3.4	Prognosis/Outcome	1[1,0]

4	Psychiatry	Yes	IN4.1	Overview	8[6,2]
			IN4.2	Diagnosis/Symptoms	8[5,3]
			IN4.3	Treatment	5[4,1]
			IN4.4	Treatment	3[1,2]
5	Urology	Yes	IN5.1	Diagnosis/Symptoms	2[2,0]
			IN5.2	Diagnosis/Symptoms	3[3,0]
			IN5.3	Prevention/Screening	4[4,0]
			IN5.4	Disease Management	2[2,0]

Procedure

Each user chose two information needs, belonging to the same or different tasks and four search engines of any type. For each information need, users had to formulate a query and submit it to the selected search engines. Following the pooling approach, each user assessed the relevance of the top-30 documents returned by each engine in a 3-graded scale (0-non relevant; 1-partially relevant and 2-totally relevant).

Participants and their choices

Forty-one undergraduate users participated in this study (27 females; 14 males) with a mean age of 27.2 years (SD=10.02). These users evaluated 9,572 documents, less than 41x2x4x30 because some queries returned less than 30 documents. Totally there were 82 sets of judged documents, one for each pair of user and information need, from which repeated documents obtained in different search engines, were excluded. As can be seen in Table 6, with the exception of one information need, all were associated with at least one user. All users chose Google as one of the four engines. In Figure 1, we present the number of users selecting each search engine.

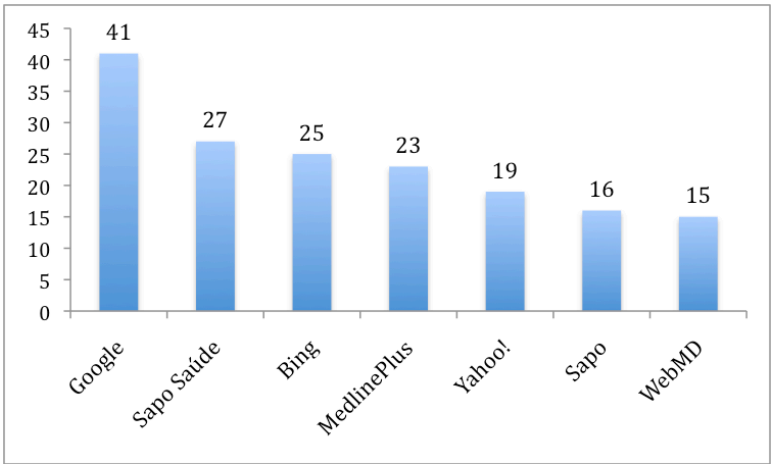

Figure 1 - Number of users selecting each search engine

Users' search behaviour was limited in the sense that they could only choose search engines from a predefined list and they had to focus on the top-30 documents. Users' querying behaviour in this experiment was already analysed in a previous work [omitted for anonymity reasons].

Evaluation measures

To evaluate and compare the engines we use binary and graded relevance measures. The first type of measures includes the Average Precision (AP), precision at 5 documents retrieved (P@5) and P@10. For computing these measures we converted the 3-graded scale into a binary one. All the partially relevant and totally relevant documents in our user study were considered relevant and the others non-relevant.

The second type of measures, recently proposed by Robertson et al. (2010), is based on the probabilistic model generalize average precision to the case of multi-graded relevance. Similarly to what has been done in the binary measures, we use the Graded Average Precision (GAP), graded precision at 5 documents retrieved (gP@5) and gP@10. The GAP and gP measures consider a user model in which the user has a binary view of relevance even when using a non-binary scale of relevance. In this model, each point of relevance in the scale has a probability g_i of being the grade from which the user considers the documents relevant. The GAP and gP@n measures are defined as:

$$GAP = \frac{\sum_{n=1}^{\infty} \frac{1}{n} \sum_{m=1}^{n} \delta_{m,n}}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j}$$

$$\delta_{m,n} = \begin{cases} \sum_{j=1}^{\min(i_m, i_n)} g_j & if \quad i_m > 0 \\ 0 & otherwise \end{cases}$$

$$gP@n = \frac{1}{n} \sum_{m=1}^{n} \frac{\sum_{j=1}^{\min(i_n, i_m)} g_j}{\sum_{j=1}^{i_n} g_j}$$

g_i – Probability that the user sets the threshold at grade i, i.e., in a relevance scale of {0..c}, he considers grades $i...c$ as relevant and the others as non-relevant.

R_i - Total number of documents in grade I for this query

i_n - Relevance grade of document at rank n

If i_n>0, document at rank n will contribute to the calculations.

More details on the calculation of these measures can be seen in the cited paper. Based on the evaluation results presented by the measure's proponents, we will use an equally balanced g_1 and g_2, i.e., $g_1=g_2=0,5$. In their experiment, considering relevant (1) and highly relevant (2) as relevant and these threshold probabilities, authors concluded that this measure is always more informative than nDCG and AP.

All measures, binary and non-binary, will be averaged over assessment cycles. An assessment cycle is composed by the set of relevance assessments of a specific user for a certain information need in one search engine. As we are comparing means, we will in fact be comparing the Mean Average Precision (MAP) and Mean Generalized Average Precision (MGAP).

We decided to use GAP and gP in our study because it is a recently proposed measure and GAP consistently outperformed nDCG and has the properties of AP that led to its predominance (Robertson et al., 2010). We decided to use binary and graded relevance measures because we want to evaluate the impact of using this new measure, comparing it to MAP, one of the most common measures. In fact we are comparing different

threshold probabilities in the model underlying GAP: $g_1=g_2=0,5$ and $g_1=1$. The last equals AP since all partially and totally relevant documents convert to relevant.

In the AP and GAP calculations we had to estimate the size of the set of relevant documents. In this sense, we considered the set of relevant documents (assessed with 1 or 2 in the relevance scale), in each pair of information need (see Table 6) and search engine, regardless of the user. It is defined as:

$$Rel(in,se) = \{doc: doc \in P(in,se) \wedge \exists\ (RJ(doc)=1 \mid RJ(doc)=2)\}$$

In this formula *in* is the information need, *se* is the search engine, *doc* is the document, *P(in,se)* is the pool of judged docs to information need *in* and search engine *se* and *RJ(doc)* is a relevance judgment made for document *doc*.

In GAP calculations, we considered the proportion of documents in *Rel(in,se)* classified with 1 and the proportion of documents classified with 2.

To prevent biases, as each assessment cycle contains at most 30 judgements, if |Rel(in,se)|>30, we only considered the existence of 30 relevant documents in MAP computation. In these cases, to MGAP, we multiply the proportion of partially relevant documents by 30 and we did the same to totally relevant documents. Without this upper limit, it would be unfair to the most selected search engines that, probably, have larger collections and only 30 judgments in assessment cycle.

4. DATA ANALYSIS

Our analysis was done in four perspectives, as depicted in Figure 2. Initially we did a global analysis with 4 goals: (1) to find if and where are the differences in the performance of each category of search engine (health and non-health) and, more specifically, on each search engine; (2) if and where are the differences in the answers to the several types of clinical questions; (3) to the different medical specialties and (4) to find if severe and non-severe conditions are associated with different global performance.

We then focused on the differences related to the types of clinical questions. Here, we investigate (5) if, in each type of search engine and in each specific search engine, there are differences in the performance for different types of clinical questions. We also studied (6) the differences that exist in each type of clinical questions (e.g.: in treatment information needs, are there differences between types of search engines and between search engines?). We followed this same strategy to analyse the differences in the medical specialties and within levels of severity (7-10).

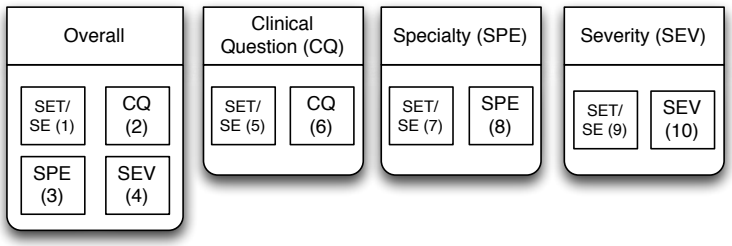

Figure 2 - Conducted analysis (SET=Search Engine Type; SE=Search Engine)

As previously stated, our analysis is based on six measures: AP, P@5, P@10, GAP, gP@5 and gP@10. This set of measures was calculated for each assessment cycle, defined by the triplet: user, information need and search engine. The mean of each of these measures was then compared between different groups using hypothesis tests. We

followed the strategy presented in Figure 3, in each measure. Whenever possible, we applied a parametric test instead of a non-parametric one due to its greater statistical power. To select the appropriate statistical test we considered the number of groups being compared. When more than two groups were being compared, we initially applied a one-way ANOVA or a Kruskal-Wallis test to detect if there were differences between the groups. If differences were found, we either applied the Tukey's test or a pairwise comparison in which we divided the α value by the total number of comparisons to minimize the type I error. These comparisons allowed us to detect where the differences are located. In comparisons between two groups we either applied the t-test or the Mann-Whitney to detect if and in what way there are differences. In all the comparisons we only considered groups with at least 5 assessment cycles in it.

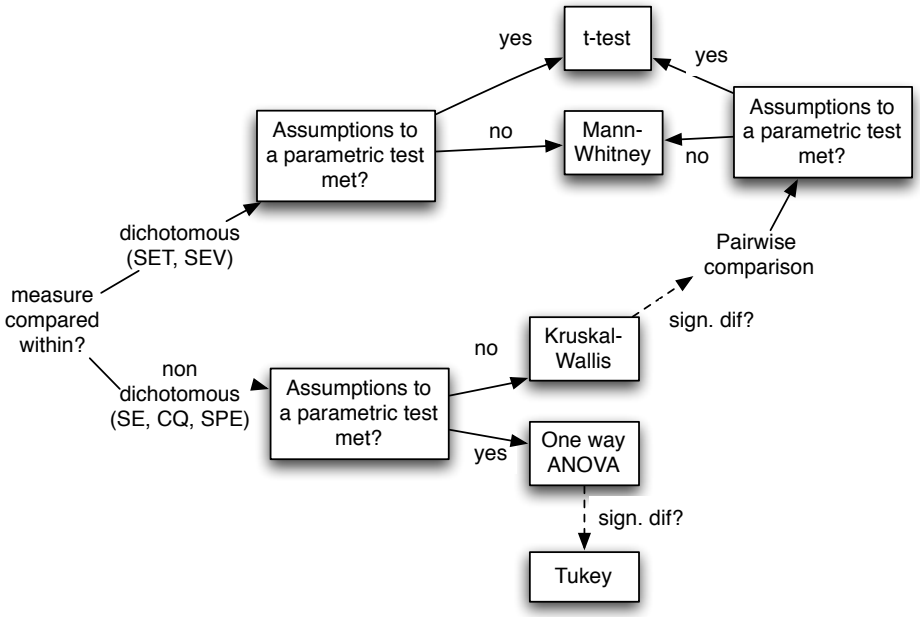

Figure 3 – Statistical strategy

As a result of the large number of hypothesis tests performed, in the next sections we will only report significant results at α=0.05 or α=0.01. Detailed results of the hypothesis tests are available at [omitted for anonymity reasons]. In the overall analysis we also present boxplots to graphically depict the GAP between groups. We chose GAP because it is an average of graded measure, therefore conveying a more stable and genuine result.

Overall analysis

In the broad analysis of differences in types of search engines, we can see that generalist search engines clearly have better performance than health-specific ones. This is not only visible in the boxplots presented in Figure 4 but also a significant difference found in all measures as indicated in Table 7.

In Figure 5, two search engines stand out, Google in a positive way and SapoSaúde in a negative way. These differences are significant in several measures and in several pairs

of engines as can be seen in Table 7. Google is significantly better than all the other engines, mainly in top-5 and top-10 measures, and SapoSaúde worst than Bing, Google, MedlinePlus and Yahoo! in the top-5 and top-10 measures. It is interesting to note that, in average measures, Google is significantly better than all the health-specific engines. Sapo is the engine with the largest statistical dispersion.

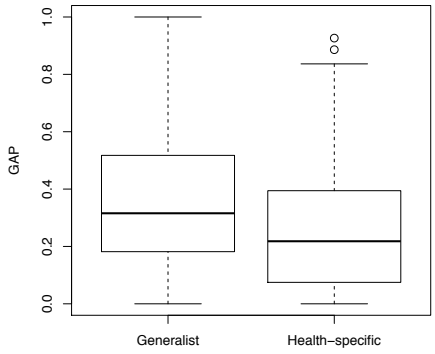

Figure 4 – GAP comparison between search engine type

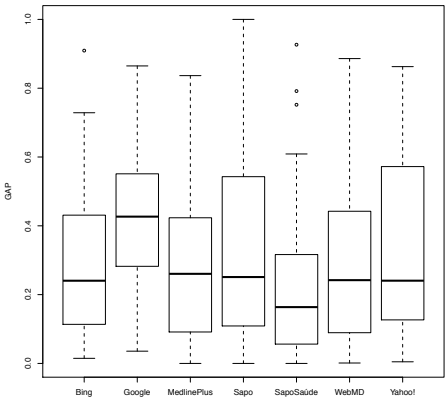

Figure 5 – GAP comparison between search engines

In the clinical question analysis, we found that precision at the top of the ranking is significantly better in the overview and diagnosis/symptoms questions than in the prevention/screening ones. As can be seen in Figure 6, differences in GAP are less evident.

Through Figure 7, we can see that Urology is the specialty with highest GAP mean and also the one with lower dispersion. Yet, there are no significant differences in GAP, only in top-5 and top-10 measures where it is clear that psychiatry is better than dermatology.

Finally, we also found that information needs associated with severe conditions have significantly higher performance than non-severe in all measures. This makes us feel there is more online information about severe health topics than non-severe ones. This is in line with what White and Horvitz (2009) conclude in their study about cyberchondria: "Web search engines have the potential to escalate medical concerns".

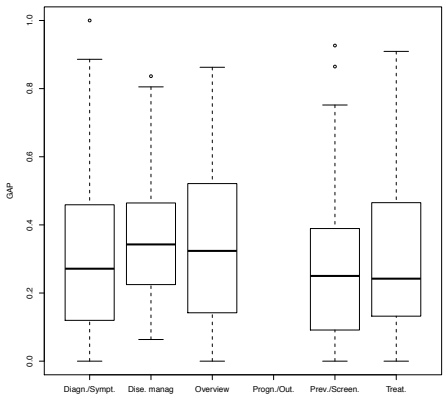

Figure 6 – GAP comparison between query types. (Not enough data for Progn./Out.)

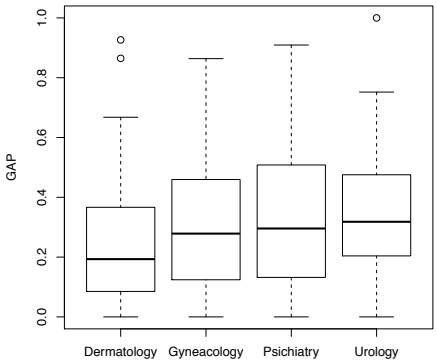

Figure 7 – GAP comparison between specialties

Table 7 – Significant differences in the overall analysis.

Comparisons	Measures		
	@5	@10	Average
Generalist>Health-specific engine	gp,p	gp,p	gap,ap
Bing>SapoSaúde	gp	gp,p	
Google>Bing Google>MedlinePlus Google>SapoSaúde	gp,p	gp,p	gap,ap
Google>Sapo	gp,p	gp,p	
Google>WebMD	p	p	gap,ap
Google>Yahoo	gp,p	gp,p	
MedlinePlus>SapoSaúde Yahoo>SapoSaúde	gp	gp	
Overview>Prevention/Screening	gp,p	p	
Diagnosis/Symptoms>Prevention/Screening	gp	p	
Gynaecology>Dermatology	gp		
Psychiatry>Dermatology	gp,p	gp,p	
Severe>Non-severe	gp,p	gp,p	gap,ap

Clinical query type analysis

Our analysis by search engine type (Table 8) shows that, in generalist search engines and the top documents, overview questions have higher precision than the prevention/screening and treatment ones. In overview, diagnosis/symptoms and prevention/screening questions, almost all measures show that generalist engines have a better precision.

Table 8 – Significant differences in the query type analysis by search engine type

Comparisons		Measures		
Where	**Which**	**@5**	**@10**	**Average**
Generalist engine	Overview>Prevention/Screening Overview>Treatment	p	gp,p	
Overview	Generalist>Health-specificSE	gp,p	gp,p	gap,ap
Diagnosis/Symptoms				
Prevention/Screening		gp,p	gp,p	ap

In Table 9 we see that, in Yahoo!, the overview questions have better precision in the top-10 documents than treatment questions. An analysis on types of clinical questions repeatedly shows that Google is better than other engines. This is more evident on the top-5 and top-10 measures. MGAP and MAP show that Google is better than SapoSaúde in the overview and diagnosis/symptoms questions. It has also a larger MAP than MedlinePlus in diagnosis/symptoms questions.

Table 9 – Significant differences in the query type analysis by search engine

Comparisons		Measures		
Where	**Which**	**@5**	**@10**	**Average**
Yahoo!	Overview>Treatment		p	
Overview	Google>SapoSaúde	gp,p	gp,p	gap,ap
Diagnosis/Symptoms	Google>SapoSaúde	gp,p	gp,p	gap,ap
	Google>MedlinePlus			ap
Prevention/Screening	Google>Sapo	p	p	
	Google>WebMD	p	gp,p	
	Google>SapoSaúde	gp	gp,p	
Treatment	Google>Yahoo!		p	

Medical specialty analysis

Table 10 shows that generalist search engines have better precision in the top documents in gynaecology questions when compared to dermatology ones. All specialties have higher top-10 measures on generalist search engines. In average, this happens in psychiatry and urology.

Table 10 – Significant differences in the medical specialty analysis by search engine type

Comparisons		Measures		
Where	**Which**	**@5**	**@10**	**Average**
Generalist engine	Gynaecology>Dermatology	gp,p	p	
Dermatology	Generalist>Health-specific engine	gp,p	gp,p	
Gynaecology				
Psychiatry			gp,p	gap,ap
Urology				

In Table 11 we see that, in MedlinePlus and WebMD, psychiatry questions have higher graded precision in the top-5 documents when compared with dermatology ones.

In this type of questions, Google surpasses 4 engines in the top-5 precision. All measures show us that Google is better than SapoSaúde in gynaecology and psychiatry questions. With the top-10 measures Google is also better than Sapo.

Table 11 – Significant differences in the medical specialty analysis by search engine

Comparisons		Measures		
Where	**Which**	**@5**	**@10**	**Average**
MedlinePlus	Psychiatry>Dermatology	gp		
WebMD	Psychiatry>Dermatology	gp		
Dermatology	Google>MedlinePlus Google>Sapo Google>SapoSaúde Google>WebMD	p		
Gynaecology	Google>MedlinePlus		p	
	Google>SapoSaúde	gp,p	gp,p	gap,ap
	Yahoo!>SapoSaúde		p	
Psychiatry	Google>Sapo	gp,p	gp,p	
	Google>SapoSaúde			gap,ap
Urology	Google>SapoSaúde		p	
	Google>Sapo		gp	

Condition severity analysis

As can be seen in Table 12, severe questions have better results in both types of engines but this is more expressive in generalist ones. The tendency of generalist engines to have better performance is also visible in both levels of severity, although more in severe ones. Does this mean that health engines have concerns on the balance of health information regarding all types of conditions?

Table 12 – Significant differences in the severity analysis by search engine type

Comparisons		Measures		
Where	**Which**	**@5**	**@10**	**Average**
Generalist engine	Severe>Non-severe	gp,p	gp,p	gap,ap
Health-specific engine			gp	
Non-severe	Generalist>Health-specific engine	gp,p	gp,p	
Severe				gap,ap

The same tendency expressed above is found on the analysis by search engine (Table 13), i.e., severe questions have better performance than non-severe ones. In Bing, Google, Sapo and WebMD, average measures are significantly higher in severe questions. In MedlinePlus, Sapo and WebMD, this superiority is also expressed in top-5 and top-10 measures. In non-severe questions, Google is better than Sapo, SapoSaúde and WebMD in top documents. In severe questions, we can also see that Google is consistently the one with better precision in pairwise comparisons and the opposite happens with SapoSaúde.

Table 13 – Significant differences in the severity analysis by search engine

Comparisons		Measures		
Where	**Which**	**@5**	**@10**	**Average**
Bing	Severe>Non-severe			gap,ap
Google				
MedlinePlus		gp,p		
Sapo		p		gap
WebMD		gp,p	gp	gap,ap
Non-severe	Google>Sapo	gp,p		
	Google>SapoSaude		gp	
	Google>WebMD	p	gp	
Severe	Google>Bing	gp,p	p	
	Bing>SapoSaúde	gp	gp,p	
	Google>MedlinePlus		gp,p	gap,ap
	Google>Sapo	gp,p	gp,p	
	Google>SapoSaúde			gap,ap
	Google>Yahoo!	p	p	gap
	MedlinePlus>SapoSaúde	gp	p	
	WebMD>SapoSaúde	gp	gp,p	
	Yahoo!>SapoSaúde			

5. DISCUSSION AND IMPLICATIONS

We compared the performance of generalist and health-specific engines satisfying health information needs. Results will be discussed next along with their implications to the user and to system development. A secondary goal of our work was to compare a recently proposed measure based on graded assessments with the traditional average precision. This comparison will be made in the end of this section.

Users' preference on Google was clear since all the participants chose Google as one of the search engines to use. American habits and preferences are similar. If we consider the proportion of sessions that starts on generalist engines (66%, as reported previously) and the latest market share of Google among them (85%, according to http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4), we can predict that 56,1% of all American health sessions start on Google. According to our results, this is a good habit since Google has shown significantly higher precision than other search engines. Differences are even more expressive in the top documents which means Google's first results page is a good place to start a health search session.

In a global perspective, generalist search engines surpass health-specific ones in precision, and this is in accordance with almost all the studies mentioned in the literature review. Yet, health-specific engines may be more balanced in the type of contents they provide in terms of severity. Indeed, although both type of engines show higher precision in severe conditions, a smaller number of significant differences is found on health-specific ones. Therefore, in order reduce the bias of the results, it might be a good practice to complement the results gathered from generalist search engines with the ones given by health-specific engines.

The higher precision obtained for severe conditions make us suspect there is more online information about severe health topics than non-severe ones, and this may raise the

problem of escalations on medical concerns. This finding alerts to the potential danger of online health information and should be considered in systems' development.

Overview clinical questions tend to have higher precision, mainly in the top results. In the complete set of search engines, they are better than the prevention/screening questions and, in generalist engines they are also better than the treatment ones. Conceptually this type of question is more comprehensive than the others and this may explain the better results. When other clinical queries have bad results, a good strategy may be their conversion to an overview type with which a user may get the specific information they want.

In the top-5 and top-10 results, gynaecology and psychiatry medical specialties have better performance than dermatology questions in the complete set of engines. This difference is more evident in the psychiatry specialty. Has the Web more and better information on this topic? Is it easier to discuss this kind of topics online? In generalist engines, only the gynaecology superiority stands.

To evaluate the relationship of our results with the topics/medical specialties popularity we have estimated the popularity of the medical condition behind each work task in two axes: number of web pages and number of searches on that topic. The number of web pages estimate was based on Google's total number of results for a query with the medical condition. On the other hand, the number of searches was estimated using on Google Trends the same expression/query. The results were then aggregated by their medical specialty and normalized. Figure 8 presents these values and also the mean of Google GAP and the mean of the overall GAP for each specialty. These two last measures were also depicted to help analyse the relation between popularity and search engines' performance. In particular, Google GAP was included because our popularity estimates were based on Google information.

In Figure 8 we also see that psychiatry and gynaecology topics are the most popular, which may help explain the significant differences mentioned above. In this figure, the Urology specialty contradicts this tendency, being an unpopular specialty with the highest GAP mean. Although this superiority in performance is not significant, this led us to analyse the correlation between popularity and GAP mean. We found a correlation of 0.34 with the number of pages and of 0.41 with the number of searches. The correlation is not high, which may imply that the search engines' performance is explained not only by topics' popularity but also by other factors like users' context.

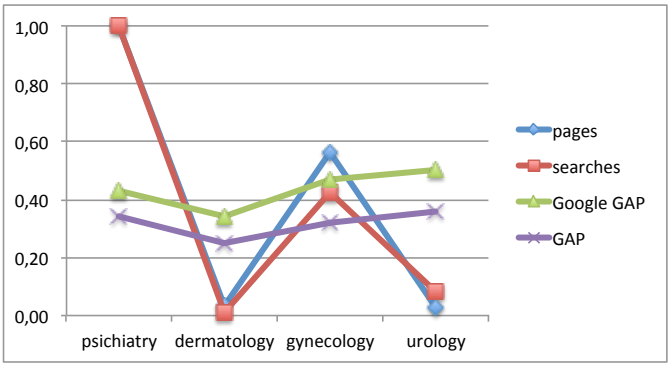

Figure 8 – Popularity of the topics' medical specialties

One of our goals was also to compare graded average precision and average precision or, in other words, to compare different threshold probabilities in the model underlying

GAP. The first threshold probabilities were defined based on the results of the measure's proponents (g_1=g_2=0.5) and the second is associated to the commonly used average precision (g_1=1, g_2=0). In Figure 9 we see that both types of measures have a very similar behaviour across search engines. The main difference lays in the magnitude of values. Generally, precision values are 0.1 higher than graded precision ones. This is expected, since the first type considers all the documents assessed with 1 and 2 relevant and, in the second, a document assessed with 1 has only a 0.5 probability of being relevant. In each type of measure, and also as expected, precision at 5 is higher than precision at 10 which, in turn, is higher than average precision.

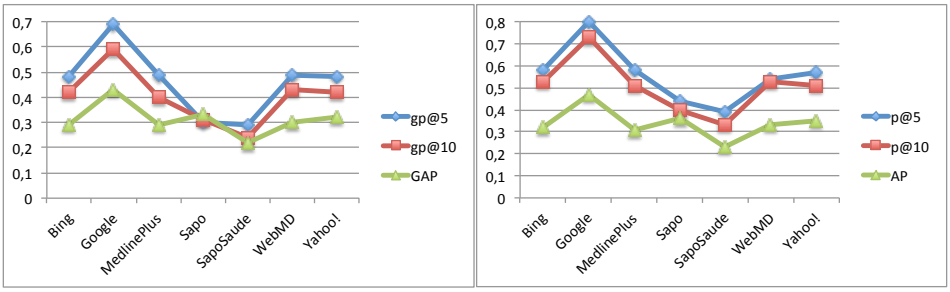

Figure 9 – Average measures in each search engine

We also analysed the significant differences found with each measure. In Figure 10 we present the number of differences found with each measure. This number tends to decrease as the number of results in the calculation increases. In GAP and AP the number of differences is smaller than on the other measures. This was expected since these measures are more aggregated and stable. They not only average but also consider more results. The exception to this trend happens with p@10 in which we found more differences than with p@5.

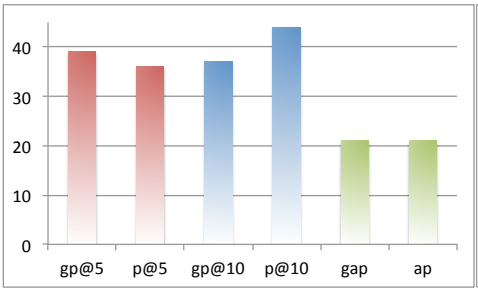

Figure 10 – Number of significant differences in each measure

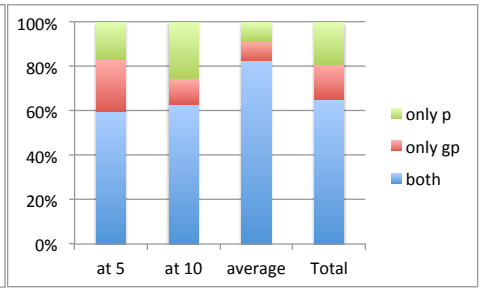

Figure 11 – Proportion of types of significant differences found in each level

In Figure 11 we present the proportion of differences found with both or only one of the measures. More than 60% of measures are significant in both types of precision (p and gp). This proportion rises to more than 80% if we use more complex measures like GAP and AP. This is in line with the previously commented stability of these measures. From this analysis we can conclude that, in evaluations that use simple measures like precision at certain rank cut-offs with graded relevance assessments, it is more critical to have an appropriate threshold definition in graded precision. In our case, we think the

first set of threshold probabilities (g_1=g_2=0.5) is more sensible and genuine because it is defined over the space of users and considers the differences between them.

6. CONCLUSION

We have conducted a user study that allowed the evaluation of seven different search engines on the health domain. Four are generalist search engines and the other are health-specific. We have compared the precision of search engines using 6 different measures in a global perspective and in specific types of information needs.

Our results show that generalist search engines surpass health-specific ones in precision. Google is users' preferred engine and it is also the one with better precision. Differences in this engine are more expressive in the top of the rank which means Google's first results page is a good place to start a health search session. To reduce the bias towards severe topics it might be a good practice to use a health-specific engine to refine results. In fact, health-specific engines seem more balanced in severity in their collections. The higher precision of severe conditions make us suspect there is more online information about severe topics than non-severe ones what may raise escalations on medical concerns.

About measures we found that complex measures like AP and GAP are less vulnerable to thresholds definition in graded precision. In evaluations using only simple measures like precision at certain rank cut-offs, it is important to have an adequate definition of these probabilities.

As future work it would be interesting to ask health experts to evaluate documents' contents and to analyse the correlation between users and experts assessments. It would also be appealing to try other threshold probabilities like g_1=0, g_2=1, in which only documents assessed with 2 (totally relevant) are considered relevant. This was an exploratory study and so, conclusions on clinical query types and medical specialties should be further studied in more focused and restricted studies.

REFERENCES

Bin, L. & Lun, K. (2001), "The retrieval effectiveness of medical information on the web". *International Journal of Medical Informatics,* Vol. 62 No. 2-3, pp. 155-163.

Borlund, P. (2003), "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems". *Information Research,* Vol. 8 No. 3.

Chu, H. & Rosenthal, M. (1996), "Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology", in *Proceedings of the American Society for Information Science Annual Meeting*, *1996*, pp. 127-135.

Eysenbach, G. & Thomson, M. (2007), "The FA4CT algorithm: a new model and tool for consumers to assess and filter health information on the Internet". *Studies in Health Technology and Informatics,* Vol. 129, pp. 142-146.

Fox, S. (2006), "Online Health Search 2006", report, Pew Internet & American Life Project, 29 October.

Fox, S. & Jones, S. (2009), "The Social Life of Health Information", report, Pew Internet & American Life Project, 11 June.

Gordon, M. & Pathak, P. (1999), "Finding information on the World Wide Web: the retrieval effectiveness of search engines". *Information Processing & Management,* Vol. 35 No. 2, pp. 141-180.

Graber, M. A., Bergus, G. R. & York, C. (1999), "Using the World Wide Web to answer clinical questions: how efficient are different methods of information retrieval?". *The Journal of Family Practice,* Vol. 48 No. 7, pp. 520-524.

Gwizdka, J. & Chignell, M. H. (1999), "Towards information retrieval measures for evaluation of web search engines", report, Interactive Media Lab, University of Toronto.

Hawking, D., Craswell, N., Bailey, P. & Griffihs, K. (2001), "Measuring Search Engine Quality". *Information Retrieval,* Vol. 4 No. 1, pp. 33-59.

Hersh, W. (2008), *Information Retrieval: A Health and Biomedical Perspective (Health Informatics)*, Springer, New York, NY, USA.

Ilic, D., Bessell, T. L., Silagy, C. A. & Green, S. (2003), "Specialized medical search-engines are no better than general search-engines in sourcing consumer information about androgen deficiency". *Human Reproduction,* Vol. 18 No. 3, pp. 557-561.

Ingwersen, P. (2009), "The User in Interactive Information Retrieval Evaluation", presentation at European Summer School in Information Retrieval (ESSIR), Padova, Italy, 2009.

Ingwersen, P. & Järvelin, K. (2005), *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*, Springer, Dordrecht, The Netherlands.

Järvelin, K. & Kekäläinen, J. (2002), "Cumulated gain-based evaluation of IR techniques". *ACM Transactions on Information Systems (ACM TOIS),* Vol. 20 No. 4, pp. 422-446.

Johnson, P., Chen, J., Eng, J., Makary, M. & Fishman, E. (2008), "A comparison of World Wide Web resources for identifying medical information". *Academic Radiology,* Vol. 15 No. 9, pp. 1165-1172.

Jones, D. & Timm, D. (2008), "Consumer Health Search Engines Comparison". *Journal of Hospital Librarianship,* Vol. 8 No. 4, pp. 418-432.

Knight, D., Holt, A. & Warren, J. (2009), "Search engines: a study of nine search engines in four categories". *Journal of Health Informatics in Developing Countries,* Vol. 3 No. 1, pp. 1-8.

Kumar, A. (2005), *Health Search Tool Evaluation.* Masters of Biomedical Informatics, Oregon Health & Science University.

Lewandowski, D. (2008), "The retrieval effectiveness of web search engines: considering results descriptions". *Journal of Documentation,* Vol. 64 No. 6, pp. 915-937.

Lopes, C. & Ribeiro, C. (2010), "Context in Health Information Retrieval: What and Where", in *Proceedings of the Fourth Workshop on Human-Computer Interaction and Information Retrieval, New Brunswick, New Jersey, USA*, *2010*.

Manning, C., Raghavan, P. & Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge.

Robertson, S., Kanoulas, E. & Yilmaz, E. (2010), "Extending average precision to graded relevance judgments", in *Proceeding of the 33rd international ACM SIGIR conference on Research*

and development in information retrieval, Geneva, Switzerland, *2010*, ACM, pp. 603-610.

Sanderson, M. (2010), "Test Collection Based Evaluation of Information Retrieval Systems". *Foundations and Trends in Information Retrieval,* Vol. 4 No. 4, pp. 247-375.

Shang, Y. & Li, L. (2002), "Precision Evaluation of Search Engines". *World Wide Web,* Vol. 5 No. 2, pp. 159-173.

Su, L. (2003a), "A comprehensive and systematic model of user evaluation of web search engines: I. theory and background". *Journal of the American Society for Information Science and Technology,* Vol. 54 No. 13, pp. 1175-1192.

Su, L. (2003b), "A comprehensive and systematic model of user evaluation of web search engines: II. an evaluation by undergraduates". *Journal of the American Society for Information Science and Technology,* Vol. 54 No. 13, pp. 1193-1223.

Tang, M. C. & Sun, Y. (2003), "Evaluation of Web-based search engines using user-effort measures". *LIBRES: Library and Information Science Research Electronic Journal,* Vol. 13 No. 2.

Tang, T., Craswell, N., Hawking, D., Griffiths, K. & Christensen, H. (2006), "Quality and relevance of domain-specific search: A case study in mental health". *Information Retrieval,* Vol. 9 No. 2, pp. 207-225.

Vaughan, L. (2004), "New measurements for search engine evaluation proposed and tested". *Information Processing and Management: an International Journal,* Vol. 40 No. 4, pp. 677-691.

White, R. & Horvitz, E. (2009), "Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search". *ACM Transactions on Information Systems (ACM TOIS),* Vol. 27 No. 4, pp. 23:1-23:37.

Wu, G. & Li, J. (1999), "Comparing Web search engine performance in searching consumer health information: evaluation and recommendations". *Bulletin of the Medical Library Association,* Vol. 87 No. 4, pp. 456-461.

Yu, H. & Kaufman, D. (2007), "A cognitive evaluation of four online search engines for answering definitional questions posed by physicians", in *Pacific Symposium on Biocomputing*, *2007*, pp. 328-339.