Context Effect on Query Formulation and Subjective Relevance in Health Searches

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

ABSTRACT

It is recognized by the Information Retrieval community that context affects the retrieval process. Query formulation and relevance assessment are stages where the user role is central. The first determines what the system will search for and the second is frequently used to evaluate how the system behaved. With a large human involvement, these stages are expected to be largely influenced by user and task characteristics. To analyze the influence of these context features on the specified stages of health information retrieval, we conducted a user study in which we collected user features through two questionnaires. User characteristics include features like age, gender, web search experience, health search experience and familiarity with the medical topic. Task features include the medical specialty, the question type, the task's clarity and the task's easiness. Besides user and task features, the relevance assessment analysis also covered features related to the query and document. We found many variables do indeed affect query formulation and relevance judgment. Some of our results question evaluations using test collections and ask for evaluation models that incorporate other kind of success measures.

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, Human Factors.

Keywords: Evaluation, Health, Relevance, User study.

1. INTRODUCTION

Several authors agree that context, often ignored, might be used to improve the retrieval process [3, 12]. Context is a loose concept and is defined in the literature in many different ways [15]. Dey & Abowd [7] present a comprehensive definition, describing context as: "any information that can be used to characterize the situation of entities (e.g. a person, a place or an object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves". Here, context is considered an interactional problem, as defined by Dourish [8]. It not only includes the environmental features surrounding the user and his activities, but also with the interaction in which he is involved. We believe context is dynamic and might change each time a new search is made, a new set of results is reviewed or a new document is viewed [10].

In the retrieval process, the interaction of the user with the system is concentrated in the formulation of the query and in the relevance assessment of the retrieved documents. With a large human involvement, these stages are expected to be largely influenced by context, as defined above. Understanding how context affects the formulation of queries can help delineate new ways, with or without the user intervention, to improve the queries as a translation of users' information needs. On the other hand, it is crucial to comprehend what factors affect relevance judgments, in which ways and how can these be incorporated in Information Retrieval (IR) systems. These factors would certainly be useful as an input to algorithms that match information needs and documents and to help IR systems move to a concept of relevance that encompasses the search context. Also, these features can be used to improve existing interfaces, either in the first stage where the user transmits the system his information need or in the latest stage, in which he accesses the retrieved documents.

There is an increasing tendency of patients, their family and friends to use the Web to search for health information [6]. The last Pew Internet report on health information [9] reveals that 61% of the american adults look online for health information. In the Internet users, this proportion rises to 83%. According to Lin and Fushman [14], this domain is extremely rich and "very well-suited for experiments in building richer models of the information seeking process".

This work intents to analyze the influence of user and task features on the formulation of queries and on the relevance assessment stages and also of query and document features on the relevance assessment stage. It will focus on health information retrieval because it is a domain with great potential in the exploration of context, it is becoming more and more common and because it is of major importance to have well-informed health consumers. The work presented here is based on a user study conducted with work tasks as

IIiX 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.

proposed by Borlund [5]. We focused on user features like age, gender, health status, web search experience, health search experience and familiarity with the topic. Regarding task features, we focused on its clarity and easiness and also on its medical specialty and clinical type (e.g. diagnosis, treatment).

This work is broader than the existing research on the influence of context features on query formulation. On the one hand it covers context features not explored before, like the health-specific ones. On the other hand, existing research is mainly focused on user expertise and type of search (e.g. exploratory, fact-finding). When compared with research that explores relevance judgments, this work is innovative because it is not based on criteria explicitly gathered from users but on implicitly gathered characteristics. Existing research is essentially based on users' explicit descriptions of what affects their relevance judgments. As users have often difficulty discussing their criteria [11], we feel implicit methods might give different insights.

In the two following sections we describe the main research done in query formulation and relevance assessment in IR. Section 4 presents the methodology underneath this study. Context influences are analyzed in two sections. Section 5 is focused on query formulation according to the query language, the use of advanced and boolean operators, the use of professional medical terminology and the number of terms. Regarding relevance, this section gives more emphasis on motivational relevance, evaluated through users self-evaluation of web search success and health search success. Section 6 does a relevance assessment analysis and is organized by categories of context features. This section focuses on situational relevance, evaluated through users relevance assessments. It also compares both types of relevance. In Section 7 we discuss the results described in the previous sections and, in Section 8, we present our conclusions and lines of future work.

2. QUERY FORMULATION IN IR

Query formulation is the process of transforming an information need into a request according to the rules of the IR system. When communicating, humans are influenced by their previous experiences and their social, organizational and cultural environment [11]. Inevitably the same happens when they formulate queries to express their needs.

Research in query formulation is usually based on analysis of log files and is traditionally more quantitative. Jansen and Pooch [13] do a good review of studies focused on web search and report that queries are often short, having only 1 or 2 terms and lack structure and language operators. Only 9% of the queries use advanced operators and only 8% use boolean operators.

Research that explores context features affecting web search is not abundant and often ignores features related to the user, the task or the concepts presented in the query [1]. In the existing studies, the most examined features are the user's expertise and the type of search.

Aula [1] conducted a user study to analyze which factors affect query formulation in web search and grouped them in three main classes: media expertise (e.g. computer, Web, search engine), domain expertise and type of search task (fact-finding, exploratory and comprehensive). In her study, media expertise is correlated with more precise and longer queries and domain expertise presumably leads to higher quality terms in queries. In fact-finding search tasks, precision is an important measure of success and, therefore, the use of precise terms or phrases is usually a good strategy. In exploratory tasks, simple queries may be enough as the goal is to obtain a general idea of the search topic and not to have high recall and precision. On the other hand, on comprehensive search tasks, a high recall is expected and a good strategy involves the use of broader terms and manual truncation.

3. RELEVANCE IN IR

The main goal of any IR system has always been the retrieval of *relevant* information. The concept of relevance is recognized as a central concern of any IR system and is related to the perceived topicality, pertinence or usefulness of documents to a particular information situation. After a large interest in the 1960s and 1970s [11], research has been stimulated again in 1990s with the work of Schamber, Eisenberg and Nilan [20].

Three insightful reviews of research on relevance are done by Saracevic in three parts [16, 18, 19], by Borlund [4] and by Ingwersen and Järvelin [11]. The section *Effects of Relevance: What Influences are Related to Relevance Judges and Judgments* in the work of Saracevic [19] is particularly pertinent as a literature review of the work reported here. For this reason, we only describe the concepts and research works most relevant to the work here presented.

3.1 Nature of relevance

Borlund [4] describes relevance as multidimensional and dynamic. It is multidimensional because it depends on the perceptions and assessments of different users and it is dynamic because it changes over time for the same user. This study only focuses on the exploration of the multidimensionality characteristic of relevance. Research in this area has been focused on the identification of the criteria used to judge the relevance of a document. In 1994, a study of Schamber [21] identifies 80 criteria as a reasonable sample of the factors used to judge relevance. In the same year, Barry [2] founds 23 criteria that were grouped in 7 categories, including the characteristics of the documents, user's previous experience, user's preferences and user's situation. The first work is a review of others' work and in the second, users are explicitly asked to explain the rationale for the relevance assessment in an interview.

3.2 Types of relevance

Relevance can be of two main types: objective/system-based relevance and subjective/user-based relevance [4]. The first is described by Saracevic [17] as the relation between a query and a document in an IR system and it is considered independent of the user, it just depends on the characteristics of the documents. IR systems are mainly based on this type of relevance because it is objective, stable and it has an easier implementation in automatic systems. This is also the concept used by the mainstream method of evaluating IR systems that incorporates a document collection, a set of requests and a set of relevance assessments, ignoring the user and his subjacent tasks.

The subjective relevance is user and context dependent and is divided by Saracevic [17] in four major categories: topical, pertinence, situational and motivational.

Topical relevance is associated with *aboutness*, this is, the relation between the topic expressed in a query and the topic expressed in a document. This type of relevance involves an assessment of the topic related to a query and a document.

Pertinence is the relation between the information need and the documents, taking into account the user's cognitive state and knowledge at the moment. This is specifically significant in health information retrieval done by consumers, in which the document's medical terminology has to be adequate to the user's knowledge to be considered relevant.

Situational relevance is expressed by the usefulness of the information objects to the user's work task.

Motivational relevance relates the user's goals and motivations with the information objects. It is expressed by the user's feeling of success and his satisfaction.

We believe that a system that incorporates features representing "persons and their interpretations/perceptions, work tasks, interaction, situations and contexts" [11] is more realistic and, therefore, in this paper, we focus on subjective types of relevance. More specifically, we will focus on situational relevance, because the study involves the user and also his interpretations of the work tasks.

3.3 Values of relevance

The scales of relevance used to judge documents are typically of two types: binary and non-binary. Binary scales are closely associated with traditional evaluation methods of IR systems using the Cranfield model. In these evaluations, documents are usually judged as *relevant* or *non-relevant*.

On the other hand, non-binary scales are more common on user-oriented IR research, becoming popular in the 1990s [11]. The number of rating values in non-binary scales differ from study to study (e.g. 11-points, 7-points, 3-points). The 3-points scale, used in this study, is the most used in IR experiments [4] and usually describes categories as: relevant, partially relevant and non-relevant.

4. METHODOLOGY

We conducted a laboratory user study with 5 work tasks based on popular questions submitted to web health support groups. The work tasks act as the context of 4 information needs that are linked to each of them. The defined work tasks are associated with the following medical specialties: gynecology, dermatology, psychiatry and urology. Moreover, 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. As an example, we transcribe one of the work tasks.

> *You are the sibling of a 5-year old child who, usually, is irritable throughout the day. There are times when you feel you can not keep up with the situation any longer but, on the other hand, you also feel sorry for her. You think she may suffer from bipolar disorder and you want to know more about this disease. For example, (T1.1) to know what characterizes the disease, (T1.2) if children can have this disease, (T1.3) how to deal with people affected by the disease and (T1.4) to know treatments for it.*

Each user chose 2 information needs (e.g. T1.1 and T3.4), regardless of the task to which they belong. Selections were distributed by the 5 tasks as follows: 20.3%, 17.6%, 17.6%, 31% and 13.5%. Then, users formulated a query for each information need which was submitted to the 4 search engines directly by the users, who chose search engines, regardless of their type, from a list of 7 search engines where 4 are generalists (Google, Bing, Yahoo! and Sapo) and 3 are health-specific (MedlinePlus, WebMD and Sapo Saúde). 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). Users were asked to, whenever possible, use the same query in every search engine. However, they were allowed to change it if the query did not return enough results or if its language needed to be adjusted to the language of the search engine's contents.

After answering an initial questionnaire, users were asked to assess relevance in a 3-graded scale of the 30 top documents returned by each of the four search engines. In the end, students also answered a final questionnaire. The initial questionnaire inquired the user on demographic data, web search experience, health seeking behavior, previous searches on the topic and knowledge on the work task. The final questionnaire included questions about the selected information needs, about the reasons that led to it and the task completion status.

Forty-one undergraduate students participated in this study (27 females; 14 males) with a mean age of 27.2 years (SD = 10.02). These students evaluated 9,572 documents, less than 41 × 2 × 4 × 30 because some queries returned less than 30 documents. The average number of years users have been searching the Web is 8.37 years (SD = 3.05), most of the students (61%) do one or two web searches a day (4 in `ws_fre` in Figure 1) and more than 80% of the students say they find what they want almost all the time (4 in `ws_suc`).

The Web is not used to search for health information by 22% of the students. As can be seen in Figure 1, the frequency of health searches (`hs_fre`) is much lower than the frequency of web searches (`ws_fre`). The majority of the students (40%) does this type of searches one or twice a month and 33% said they did it one or two times a year. In these searches, users feel less successful (`hs_suc`) than in general web searches. Globally, students consider they have a good health condition (`hstat` in Figure 1).

Only 25% of the selected information needs were about a previously searched topic. In a global perspective, as can be seen in variables `clar`, `comp` and `fam` of Figure 1, students found the tasks clear, moderately complex and were somehow familiar with the topic.

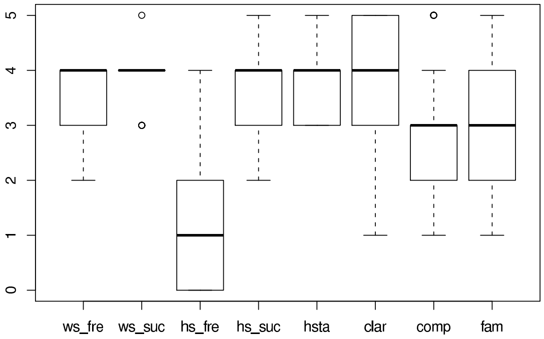

Figure 1: Distributions of ordinal variables. Variables' descriptions and scales in Tables 1 and 8.

Table 1: Context features

Dimension	Feature	Description	Scale	Values
user	age	-	ratio	-
	gender	-	nominal	female, male
web search	ws_freq	frequency	ordinal	1 (twice a year) to 5 (more than twice a day)
	ws_success	success rate	ordinal	1 (never find) to 5 (always find)
	ws_years	years of experience	ratio	-
health search	hs_freq	frequency	ordinal	1 (twice a year) to 5 (more than twice a day)
	hs_success	success rate	ordinal	1 (never find) to 5 (always find)
	hs_webuse	Web use for these searches?	nominal	no, yes
topic's familiarity	familiarity	self-evaluation of familiarity	ordinal	1 (not familiar) to 5 (familiar)
	prev_search	previous searches	nominal	no, yes
task	clarity	-	ordinal	1 (not clear) to 5 (clear)
	easiness	-	ordinal	1 (difficult) to 5 (easy)
	qtype	question type	nominal	overview (o), disease management (dm), treatment (t), prevention/screening (p/s), prognosis/outcome (p/o), diagnosis/symptoms (d/s)
	specialty	medical specialty	nominal	psiquiatry (p), dermatology (d), gynecology (g), urology (u)

5. QUERY ANALYSIS

Queries formulated by the users were analyzed in four perspectives: the language of the query terms, the use of advanced and boolean operators, the use of technical medical terms and the number of terms. The language of the query was manually labeled and the use of technical medical terms was identified based on a multilingual glossary of technical and lay medical terms[1]. The analysis was done according to the dimensions and context variables presented in Table 1.

The language, use of advanced and boolean operators and use of technical terms are all nominal variables. Therefore, we followed the strategy presented in Figure 2 in these three dimensions. We have compared the distributions of Table 1's variables in the groups defined by the above variables (e.g. portuguese and english in the language variable). With the one-tailed test in nominal and dichotomous variables, we were able to detect the direction of the differences (e.g. higher or lower).

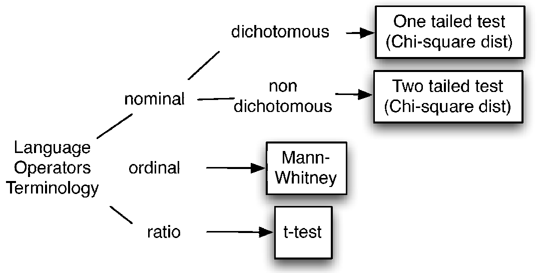

Figure 2: Statistical analysis of the language, operators and terminology variables.

The strategy to analyze the impact of context features on the number of terms is presented in Figure 3. It is different because the number of terms analysis is a ratio variable.

We have compared the average number of terms in the groups defined by nominal and ordinal variables and have analyzed its correlation with ratio variables. We have applied the Kruskal-Wallis test instead of the Anova test because the variances were not homogeneous. When we found significant differences with the Kruskal-Wallis test, we also did a pairwise comparison in which we have divided the α value by the total number of comparisons.

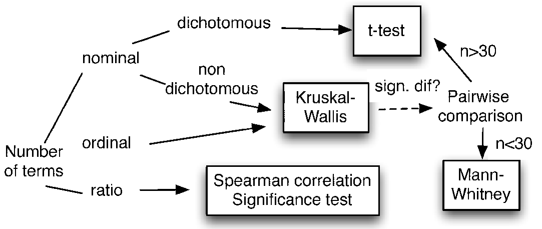

Figure 3: Statistical analysis of the number of terms variable.

5.1 Global analysis

In the conducted experiment, users issued a total of 155 different queries. User's first language, Portuguese, was used in 76% of the search sessions and English in all other sessions. Each user has done 8 search sessions, 4 to each information need. A deeper analysis shows us that all search sessions in Medline and WebMD were made in English. In Yahoo, 92% of the search sessions were made in Portuguese and, in all other search engines, Portuguese was the preferred language. This suggests that, in most cases, the use of the english language might not have been a user's choice but an imposition of the selected search engine. Only 17% of the queries used advanced or boolean operators and the average number of terms was 3.78 (SD = 2.01). The majority of the search sessions are associated with 2 (19%), 3 (37%) or 4 terms (19%). Only 3% of the queries used medical technical terms. The proportion of structured queries is similar to the one reported by Jansen and Pooch [13] and the average number of terms is slightly superior.

5.2 Language

The global analysis of the language used in queries showed that users tend to search in Portuguese, only showing a different behavior when using search engines with contents in other languages. Yet, we decided to further analyze the influence of context features on the choice of language because, in Yahoo!, some users opted for the english language.

In the column *Language* of Table 2, we can see that female users have a higher proportion of search sessions in English than the male users. Through the information of Table 3

[1] Available at: `http://users.ugent.be/~rvdstich/eugloss/welcome.html`

Table 2: Context effects of nominal variables: Chi-square test results. *p<.05; **p<.01. Question mark represents a Chi-square approximation that may be incorrect. Proportions as $p_{row}(column)$.

Var	Language	Operators	Terminology
gender	$p_f(en) > p_m(en)$ $\chi^2(1)$=12.68 p=0.00** (>)	$p_f(y) < p_m(y)$ $\chi^2(1)$= 0.26 p=0.31 (<)	$p_f(y) > p_m(y)$ $\chi^2(1)$= 1.19? p=0.14 (>)
hs_wuse	$p_n(en) < p_y(en)$ $\chi^2(1)$=1.05 p=0.15 (<)	$p_n(y) > p_y(y)$ $\chi^2(1)$=0.33 p=0.28 (>)	$p_n(y) > p_y(y)$ $\chi^2(1)$=4.78? p=0.01* (>)
prev_se.	$p_n(en) < p_y(en)$ $\chi^2(1)$=2.46 p=0.06 (<)	$p_n(y) < p_y(y)$ $\chi^2(1)$=16.19 p=0.00** (<)	$p_n(y) < p_y(y)$ $\chi^2(1)$=13.98? p=0.00** (<)
qtype	$\chi^2(5)$=2.24? p=0.81	$\chi^2(5)$=10.95? p=0.05	$\chi^2(5)$=7.53? p=0.18
specia.	$\chi^2(3)$=5.17 p=0.16	$\chi^2(3)$=35.66 p=0.00**	$\chi^2(3)$=4.38? p=0.22

Table 3: Context effects of ordinal variables: median and Mann-Whitney U test results. *p<.05; **p<.01. Signs > and < indicate one-tailed tests.

Var	Language	Operators	Terminology

clarity	EN: 5, PT: 4
	U=10375.5
	p=0.00**(>)
		N: 4, Y: 5
		U=2929.5
		p=0.00**(<)
			N: 4, Y: 5
			U=488
			p=0.03*(<)

easiness	EN: 3, PT: 3
	U=8219.5
	p=0.28(<)
		N: 3, Y: 2
		U=6307.5
		p=0.00**(>)
			N: 3, Y: 2
			U=1086
			p=0.13(>)

familiarity	EN: 3, PT: 3
	U=10039.5
	p=0.00**(>)
		N: 3, Y: 3
		U=4521.5
		p=0.28(<)
			N: 3, Y: 4
			U=378
			p=0.00**(<)

hs_freq	EN: 1, PT: 1
	U=6311.5
	p=0.25 (<)
		N: 1, Y: 1
		U=2983.5
		p=0.11(<)
			N: 1, Y: 0
			U=1006
			p=0.03*(>)

hs_success	EN: 3, PT: 4
	U=6223.5
	p=0.24(<)
		N: 4, Y: 4
		U=2335.5
		p=0.00**(<)
			N: 4, Y: 5
			U=222
			p=0.00**(<)

ws_freq	EN: 4, PT: 4
	U=6831.5
	p=0.06(>)
		N:4, Y: 4
		U=2971.5
		p=0.04*(<)
			N: 4, Y: 4
			U=438
			p=0.03*(<)

ws_success	EN: 4, PT: 4
	U=6351.5
	p=0.2(>)
		N:4, Y: 4
		U=2551.5
		p=0.03*(<)
			N: 4, Y: 4
			U=630
			p=0.21(<)

we conclude that the task's topic familiarity and the task's clarity are superior in english queries.

We detected that, in users that use the Web more often to search for information (`ws_freq`), there is a growing tendency to use English but this difference is not significant.

In the initial questionnaire users were inquired about their preferred language in web searches. Although this is a variable that is not explored in this study, we were curious to know if a systematic use of English leads to more successful web searches (`ws_success`). We found that every user that always find what he looks for (5 in `ws_success`), routinely use English in their web searches. However there was no statistical evidence of this (Mann-Whitney U=6912, p=0.06).

5.3 Advanced and boolean operators

There is statistical evidence to conclude the use of advanced or boolean operators is done more often by older users (Table 4). In Table 3 we can see that users that don't use advanced and boolean operators use the Web less often to conduct web searches and have a smaller web search success rate. This suggests that, as the experience in web search increases, users apply more structured queries, an habit that is associated with a higher rate of success. The same habit also affects positively the success rate of health searches.

Table 4: Context effects of ratio variables: mean (sd) and t-test result. *p<.05; **p<.01

Var	Language	Operators	Terminology

age	EN:27.52(9.26)
	PT:27.19(10.14)
	t(138.01)=0.25
	p=0.8
		N:26.13(8.76)
		Y:35.03(13.21)
		t(36.28)=-3.74
		p=0.00**
			N:27.3(9.98)
			Y:26.67(5.16)
			t(5.93)=0.28
			p=0.78

ws_ye.	EN:8.01(3.17)
	PT:8.51(2.93)
	t(118.54)=-1.15
	p=0.25
		N:8.54(2.75)
		Y:7.27(4.22)
		t(36.13)=1.67
		p=0.10
			N:8.34(3.03)
			Y:10(0)
			t(249)=-8.69
			p=0.00**

Users that have made previous searches on the topic use more advanced and boolean operators (Table 2). There is also evidence to state there is an association between the use of operators and the medical specialty. Structured queries are associated with a higher proportion of gynecology tasks (43%) and, in urology, all queries were simple. In Table 3, we see that structured queries are associated with more clear and difficult tasks.

5.4 Use of technical medical terms

Since only five queries, formulated by two users, employed technical medical terminology, results reported in this section do not have the same statistical strength, particularly in the Chi-square tests where the high number of cells with expected values lower than 5 amplifies the test value. When compared to the familiarity and task's variables, user, web search and health search variables have even less statistical meaning. Being aware of this situation, we still decided to present the results of our analysis as these may lead to new research hypothesis that may be studied later.

The reduced number of queries with professional terminology is, by itself, an indicator of its lack of use in information retrieval by health consumers. This reality might be explored in relevance feedback techniques provided that the terminology used in the results is adequate to the users' proficiency.

Results presented in Tables 3 and 4 show that the use of professional terms might be an habit more associated with users with longer experience on web search (`ws_years`) and a higher frequency of web searches (`ws_freq`). Contrary to our expectations, results show that the use of technical terms might be related to a smaller frequency of health searches (`hs_freq`). Results also suggest the use of professional terminology might be associated with more successful health searches (`hs_success`). In Table 3 we can see that queries with technical terms are associated with more familiar (`prev_search`, `familiarity`) and clear tasks.

5.5 Number of terms

To analyze the effects of age on the query number of terms, we have calculated the Spearman correlation coefficient, obtaining a low correlation of ρ=0.16, p<0.01**. Although age does not have a great influence on the number of terms, the gender does. As can be seen in Table 5, females use more terms per query.

The Spearman correlation between years of experience in web search and number of terms used in a query (ρ= - 0.29, p<0.01**) points out an inverse relationship with low expression and suggests that, as the number of years of ex-

Table 5: Context effects of nominal variables on the number of terms. *p<.05; **p<.01. KW stands for Kruskal-Wallis.

Var	Mean (sd)	Test	p-value
gender	F: 3.90 (2.21) M: 3.33(1.23)	t(224.30)= 2.58	p=0.00** (>)
hs_webuse	N: 3.18 (0.85) Y: 3.89(2.19)	t(229.22)= -3.72	p=0.00** (<)
prev_search	N: 3.31 (1.42) Y: 4.5(2.81)	t(88.66)= -3.53	p=0.00** (<)
qtype	O: 2.82 (0.95) DM: 3.94 (2.05) T: 2.64 (0.77) P/S: 4.82 (2.45) P/O: 6.00 (1.15) D/S: 3.98 (1.94)	KW $\chi^2(5)$= 75.20	p=0.00**
specialty	P: 2.71 (0.93) D: 4.83 (2.65) G: 4.83 (2.07) U: 4.41 (1.53)	KW $\chi^2(3)$= 113.11	p=0.00**

perience in web search increases, the number of terms gets smaller. The means presented in Table 6 show that users that search the Web more frequently have a tendency to formulate longer queries. However, in the pairwise comparison (Table 7), the only significant difference lays in the comparison of the 2^{nd} level of frequency and the 3^{rd}, in which the first has a lower median. The means of web search success (`ws_success`) made us suspect the use of more terms per query could lead to higher success rates, but differences found are not statistically significant.

Table 6: Context effects of ordinal variables on the number of terms. *p<.05; **p<.01.

Var	Mean (sd)	Kruskal-Wallis	p-value
clarity	1: 2.50 (0.58) 2: 2.75 (1.06) 3: 3.90 (1.72) 4: 2.86 (1.11) 5: 3.96 (2.42)	KW $\chi^2(4)$= 24.65	p=0.00**
easiness	1: 3.00 (1.41) 2: 3.56 (1.79) 3: 4.17 (2.21) 4: 2.41 (0.71) 5: 2.75 (1.07)	KW $\chi^2(4)$= 39.31	p=0.00**
familiarity	1: 3.25 (1.42) 2: 2.83 (0.93) 3: 4.23 (2.53) 4: 3.91 (1.99) 5: 3.47 (1.29)	KW $\chi^2(4)$= 18.93	p=0.00**
hs_freq	1: 3.44 (1.17) 2: 3.67 (1.40) 3: 3.10 (1.93) 5: 6.87 (3.55)	KW $\chi^2(3)$= 35.00	p=0.00**
hs_success	2: 3.12 (0.61) 3: 3.28 (1.37) 4: 4.01 (2.06) 5: 4.47 (3.17)	KW $\chi^2(3)$= 5.54	p=0.14
ws_freq	2: 2.83 (0.70) 3: 3.85 (1.37) 4: 3.84 (2.35)	KW $\chi^2(2)$= 9.06	p=0.01*
ws_success	3: 3.02 (0.89) 4: 3.69 (1.78) 5: 3.50 (1.77)	KW $\chi^2(2)$= 2.47	p=0.29

Table 7: Context effects on the number of query terms. Significant differences found in multiple comparisons. P-value divided by the number of tests performed. MW stands for Mann-Whitney.

Var	Difference	Test value	p-value
clarity	3>2	MW U=473.5	p<0.05/10
	3>4	t(161.98)=4.71	p<0.01/10
	5>4	t(167.77)=-4.13	p<0.01/10
easiness	2>4	t(125.67)=5.33	p<0.01/10
	3>4	t(149.06)=7.59	p<0.01/10
	3>5	MW U=2183.5	p<0.01/10
familiarity	2<3	t(118.24)=-4.89	p<0.01/10
	2<4	t(91.51)=-4.07	p<0.01/10
hs_freq	5>1	MW U=386	p<0.01/6
	5>2	MW U=515.5	p<0.01/6
	5>3	MW U=163.5	p<0.01/6
qtype	O<P/S	t(98.31)=-6.6	p<0.01/15
	O<P/O	MW U=4	p<0.01/15
	O<D/S	t(159.31)=-5.25	p<0.01/15
	T<P/S	t(94.31)=-7.27	p<0.01/15
	T<P/O	MW U=0	p<0.01/15
	T<D/S	t(148.49)=-6.18	p<0.01/15
specialty	P<D	t(64.35)=-6.08	p<0.01/6
	P<G	t(67.86)=-7.69	p<0.01/6
	P<U	t(51.75)=-7.04	p<0.01/6
ws_freq	2<3	MW U=488	p<0.01/3

There is statistical evidence to state that who uses the Web to conduct health searches, employ more terms per query. In these users, the ones that do health searches more often tend to user more terms than occasional health searchers. In fact, after the pairwise comparison, we found that the highest frequency (5) of health searches in the Web has a statistically higher median than all the other frequencies. Just like what happens in web search success, the descriptive analysis of health search success make us suppose that longer queries have higher health success rates. However, these differences are not significant.

As can be seen in Table 5, users with previous searches on the topic use more terms per query. The same happens when users are more familiar with the topic. In fact, we found that the 2^{nd} level of familiarity uses less terms than the 3^{rd} and 4^{th} levels (Table 7).

The distribution of query terms changes with medical specialties and also with query types (Table 5). Further analysis (Table 7), allowed us to conclude that the number of terms in psychiatry (P) is smaller than in all other specialties. In the query type, we found statistical evidence to say that Overview (O) and Treatment (T) questions have, in average, less terms than the Prevention/Screening (P/S), Prognosis/Outcome (P/O) and Diagnosis/Symptoms (D/S).

If the 3^{rd} level was excluded from the clarity variable, we would conclude that clarity was associated with a higher number of terms. With statistical meaning, we observe that level 3 uses more terms than level 2 and 4 and that level 5 uses more terms than level 4. Regarding the easiness of task, results show that more complex tasks are associated with longer queries. In fact, the highest levels of easiness have less terms than the 2^{nd} and 3^{rd} levels.

6. RELEVANCE JUDGMENTS ANALYSIS

In the analysis of the effects of context features on relevance judgments we have considered an additional set of variables, three in existing dimensions and the others on new dimensions. These variables are presented in Table 8.

The data analyzed in this section consists of 9572 relevance judgments. The majority of the judgements classify documents as non-relevant (58%), 26% as partially relevant and 17% as totally relevant. Distinguishing levels of relevance 1 and 2 in a scale of 0 (non-relevant), 1 (partially relevant) and 2 (totally relevant) was one of the main diffi-

Table 8: Additional context features to the relevance judgment analysis

Dim	Featu.	Description	Scale	Values
User	hstatus	Health status self-evaluation.	ordinal	1(not healthy) to 5(healthy)
Health search	usual-engine	Is this engine typically used?	nominal	no, yes
Task	task-stat	completion status	ordinal	1(failure) to 5(success)
Query	med-terms	use of medical terminology?	nominal	no, yes
Query	nterms	number of terms	ratio	-
Query	qlang	query language	nominal	EN, PT
Query	qadv	advanced or boolean operators	nominal	no, yes
Document	doc-rank	position in the ranking	ordinal	-
Document	doc-type	file type	nominal	doc, html, pdf, ppt, swf
Document	snippet	snippet length	ratio	-
Document	title	title length	ratio	-

culties felt and explicitly pointed by the users of this study. The presence of a highest peak on the non-relevance side is in accordance with what Saracevic [19] reports.

Our analysis followed the strategy explicit in Figure 4. On nominal and ordinal variables (Tables 9 and 10) we compared the median of relevance in each group of the variable. In ratio variables (Table 11) we compared the mean of the variable (e.g. age) in the three levels of relevance. We have applied the Kruskal-Wallis test instead of the Anova test because the variances were not homogeneous.

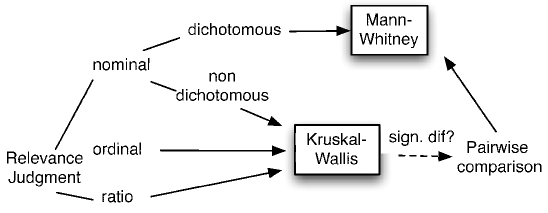

Figure 4: Relevance statistical analysis.

Table 9: Context effects of nominal dichotomous variables on relevance. *p<.05; **p<.01. MW are the initials of Mann-Whitney. All medians are 0, except the one on *usualengine* = *yes* that is 1.

Var	Test	p-value
gender	MW U= 5504548	p=0.00** (F<M)
hs_webuse	MW U= 4420862	p=0.00** (no<yes)
medterms	MW U= 1068658	p=0.00** (no>yes)
prev_search	MW U= 7420967	p=0.00** (no>yes)
qadv	MW U= 5911562	p=0.00** (no<yes)
qlang	MW U= 8229611	p=0.03* (en<pt)
usualengine	MW U=6951394	p=0.00** (no<yes)

6.1 User

In Table 11 we can see the average level of relevance increases with the age. With further analysis we verified that the average age of users in relevance 0 is lower than in relevance 1 and 2 (Table 12). These results make us conclude that younger students tend to classify documents as non-relevant more often. This raises the following question: "Do older users find documents more relevant?". Or is health information more meaningful to older students who are more sensitive to health searches and, therefore, more careful in their analysis?

Table 10: Context effects of nominal and non-dichotomous variables and ordinal variables on relevance. *p<.05; **p<.01.

Var	Kruskal-Wallis	p-value
clarity	KW $\chi^2(4)$= 39.90	p=0.00**
docrank	KW $\chi^2(2)$= 286.46	p=0.01**
doctype	KW $\chi^2(4)$= 10.18	p=0.03*
easiness	KW $\chi^2(4)$= 25.82	p=0.00**
familiarity	KW $\chi^2(4)$=25.47	p=0.00**
hs_freq	KW $\chi^2(3)$= 48.85	p=0.00**
hs_success	KW $\chi^2(3)$= 105.52	p=0.00**
hstatus	KW $\chi^2(2)$= 14.12	p=0.00**
qtype	KW $\chi^2(5)$= 85.13	p=0.00**
taskstat	KW $\chi^2(4)$= 81.96	p=0.00**
specialty	KW $\chi^2(3)$= 70.31	p=0.00**
ws_freq	KW $\chi^2(3)$= 5.87	p=0.05
ws_success	KW $\chi^2(2)$=61.56	p=0.00**

Table 11: Context effects of ratio variables on relevance. *p<.05; **p<.01.

Var	Mean (sd)	Kruskal-Wallis	p-value
age	0: 26.8 (9.27) 1: 27.86 (10.79) 2: 27.88 (10.25)	KW $\chi^2(2)$= 30.44	p=0.00**
nterms	0: 3.80 (1.99) 1: 3.65 (1.88) 2: 3.52 (1.89)	KW $\chi^2(2)$= 28.17	p=0.00**
snippet	0:105.3 (278.62) 1: 102 (75.55) 2: 108 (85.83)	KW $\chi^2(2)$= 25.15	p=0.00**
title	0:77.21 (24.93) 1: 77.49 (23.47) 2: 73.93 (23.95)	KW $\chi^2(2)$= 28.83	p=0.00**
ws_years	0: 8.05 (2.90) 1: 8.76 (3.11) 2: 8.61 (3.05)	KW $\chi^2(2)$= 85.71	p=0.00**

As seen in Table 9, male users judge documents with higher values of relevance.

In user's health status we detected significant differences on the average relevance assessed by healthier users (5th level in hstatus) and by users with the 3rd and 4th levels: 5<3 and 5<4. This suggests that healthier people judge documents with lower relevance scores. In this question no one answered the 1st and 2nd option. This result agrees with the hypothesis we raised when analyzing the age. Are healthier students less prone to health searches and have less motivation to analyze the documents in depth?

6.2 Web search experience

We found users with less years of web search experience tend to rate documents more often with 0 than with 1.

In the frequency of web searches (`ws_freq`) we found no significant differences but we detected differences in the web search success rate (`ws_success`). Not surprisingly, we found that users that feel they find everything (5 in `ws_success`) find documents more relevant: 5>3 and 5>4. Also, and not expected, we found that 3>4, this is, users that consider to have median success (3 in `ws_success`) rate relevance higher than users with 4 in `ws_success`. No user considered to have the lowest levels (1 and 2) of web search success.

Table 12: Relevance judgment analysis. Significant differences found in multiple comparisons. P-value divided by the number of tests performed. Values in *Differences* regard relevance levels in ratio variables and variable's groups in the remaining cases.

Var	Difference	Mann-Whitney	p-value
age	0<1	U= 3471313	p<0.01/3
	0<2	U= 2558583	p<0.01/3
clarity	3>4	U= 2842290	p<0.01/10
	3>5	U= 5120817	p<0.01/10
docrank	0>1	U= 7769394	p<0.01/3
	0>2	U= 5555828	p<0.01/3
	1>2	U= 2174968	p<0.01/3
doctype	pdf>html	U= 1925281	p<0.05/10
easiness	1<3	U= 396414	p<0.05/10
	1<4	U= 96912.5	p<0.01/10
	1<5	U= 67709.5	p<0.01/10
	2<3	U= 5542021	p<0.05/10
	2<4	U= 1352704	p<0.01/10
familiarity	4<1	U= 723743	p<0.05/10
	4<2	U= 2475447	p<0.01/10
	4<3	U= 2827383	p<0.01/10
hs_freq	1>2	U= 3439556	p<0.01/6
	1>3	U= 1485523	p<0.01/6
	1>5	U= 941181	p<0.01/6
	5<2	U= 1049677	p<0.05/6
	5<3	U= 459028	p<0.05/6
hs_success	5<2	U= 255937.5	p<0.01/6
	5<3	U= 1589742	p<0.01/6
	5<4	U= 1774288	p<0.01/6
	3>2	U= 606714	p<0.05/6
	3>4	U= 4798980	p<0.05/6
hstatus	5<3	U= 1760352	p<0.05/3
	5<4	U= 3223237	p<0.01/3
nterms	0>1	U= 6975251	p<0.05/3
	0>2	U= 4805382	p<0.01/3
	2<1	U= 2065639	p<0.05/3
qtype	P/S<O	U=2667458	p<0.01/15
	P/S<DM	U=536497	p<0.01/15
	P/S<T	U=1963428	p<0.01/15
	P/S<D/S	U=3125707	p<0.01/15
	P/O<O	U=100631	p<0.01/15
	P/O<DM	U=20288.5	p<0.01/15
	P/O<T	U=74228.5	p<0.01/15
	P/O<D/S	U=85668	p<0.01/15
snippet	0>1	U=7138175	p<0.01/3
specialty	P>D	U= 4671748	p<0.01/6
	P>G	U= 4328250	p<0.05/6
	P>U	U= 3250727	p<0.05/6
	D<G	U=1418530	p<0.01/6
	D<U	U=1067578	p<0.01/6
taskstat	1>2	U= 50724.5	p<0.01/10
	1>3	U= 226135	p<0.05/10
	1>5	U= 86700	p<0.01/10
	3>5	U= 2182096	p<0.01/10
	4>5	U= 4409888	p<0.01/10
	2<3	U= 1052153	p<0.01/10
	2<4	U=1004460	p<0.01/10
title	2<0	U=4809854	p<0.01/3
	2<1	U=2148352	p<0.01/3
ws_success	5>3	U= 100818.5	p<0.01/3
	5>4	U= 468468.5	p<0.01/3
	3>4	U= 3584470	p<0.01/3
ws_years	0<1	U= 2500464	p<0.01/3

6.3 Health search experience

As can be seen in Table 9, users that usually conduct health searches on the Web (`hs_webuse`) tend to rate relevance higher than the ones that don't use the Web for this purpose.

In Table 10 we can see there are significant differences in the levels of health searches' frequency and health search success rate. In health search frequency, by Table 12, we conclude that the lowest frequency in health searches is associated with higher levels of relevance and the opposite with the highest levels of frequency in health web searches. This suggests that, as the frequency of health searches rises, the relevance criterion becomes more strict.

Regarding the health search success rate, nobody answered the option 1. Surprisingly, we found that the highest level of success (5 in `hs_success`) is associated with lowest levels of relevance and that the median level of success (3 in `hs_success`) is associated with the highest levels of relevance: 5<2, 5<3, 5< 4, 3>2 and 3>4.

We also concluded that relevance is significantly higher in search engines that users typically use in their own health searches. This suggests habit leads to trust in the search engine.

6.4 Familiarity with the topic

The data in Table 9 let us see that users who have done previous searches on the topic (`prev_search`) tend to rate relevance lower than the others. This might be explained by more demanding needs. In Table 10 we see there are significant differences between the groups of self-evaluation of familiarity with the topic (`familiarity`). Further analysis allowed us to conclude that the highest level of familiarity is associated with the lowest relevance. This corroborates our suspicions based on `prev_search`.

6.5 Task

Analyzing Tables 9 and 10 we see there are significant differences between the groups of all the variables in this dimension: specialty, question type, clarity and easiness of the task. The specific differences will be described next.

In terms of clarity we found that the average clear tasks (3 in `clarity`) have higher relevance rates than the tasks classified with 4 and 5. In the clarity aspect, a clear pattern does not emerge.

The more difficult tasks have lower relevance scores. As expected, we found that tasks with the lowest rate of easiness (1 and 2 in `easiness`) have lower relevance scores then tasks with `easiness` 3, 4 and 5.

Regarding the question type we could find out that the Prevention/Screening and the Prognosis/Outcome categories, compared with all other types of questions, have the lowest relevance scores.

We can also verify that the psychiatry specialty is associated with the higher levels of relevance. Moreover, we found that the dermatology medical specialty is associated with lowest levels of relevance.

6.6 Query

In the query dimension we noticed (Table 9) that relevance is significantly higher when queries use advanced operators, when they use portuguese terms and when they use lay terms instead of technical ones. The last result contradicts Section 5.4 finding, in which we concluded that, according to hs_freq, the use of professional terminology was associated with more successful health searches. This happens because the relevance evaluated by hs_freq is motivational and differs from the situational relevance that is being studied in Section 6. In fact, we have already noticed in Section 6.3 that motivational relevance is not consistent with the situational one. It is also important to note that the use of lay terms may result in a set of retrieved documents with a language more adjusted to the health consumer and, therefore, in a result set with greater situational relevance.

The means presented in Table 11 show the number of terms decreases as relevance increases. We found significant differences in the number of terms' distributions in relevance levels. More precisely, we confirmed our suspicion, this is, relevance 0 has the largest median of terms and level 2 has the lowest median of terms: 0>1, 0>2 and 2<1.

6.7 Document

As expected, relevance decreases with the position of the document in the ranking. This tendency can be seen in the means presented in Table 11 and in the pairwise comparison.

We found differences in the relevance associated with different types of documents (Table 9) where pdf documents have higher relevance than html documents.

Analyzing if the title and snippet sizes had any influence on relevance judgments, we found that there are differences on the distributions of each of these variables in the relevance levels (Table 11). Further analysis let us conclude that non-relevant documents have longer snippets than partially relevant documents and that documents classified as totally relevant have shorter titles than non-relevant or partially relevant documents. Although title and snippet lengths may influence the decision of accessing a document, they don't seem to have impact on the assessment of relevance.

6.8 Situational versus motivational relevance

Besides exploring how do variables in Tables 1 and 8 affect relevance judgments, we also wanted to study the relationship between the situational relevance given by the relevance judgments and the motivational relevance given by the task completion status (taskstat) as perceived by the user.

In Table 10 we can see the situational relevance differs in levels of motivational relevance. With further analysis we conclude that users with a smaller feeling of success rated higher relevance scores, except in case of the level 2 that has smaller relevance scores that level 3 and 4. Although not statistically significant, this is confirmed by a negative Spearman correlation between both types of relevances (ρ=-0.02, p=0.14). This result was a surprise and is discussed in the following section.

7. DISCUSSION OF RESULTS

Based on the results presented in the previous sections, we will now discuss the main results and raise hypothesis.

Users express their queries in English less than we expected and they do it mainly because some search engines have their collections in English. Females tend to use more english terms or to select more often search engines with english content. Even though we found that portuguese queries had best situational relevance, we think this conclusion was affected by the low english proficiency of the users.

Results suggest that, as the experience in web search increases, users apply more structured queries and this is associated with a higher rate of success, motivational and situational. Also, users with higher health search success rate and users with previous searches on the topic tend to use more advanced operators.

We confirmed our hypothesis and noticed that professional terminology is seldom used by health consumers. The small number of queries with medical terminology does not allow a reliable statistical analysis. However we found tendencies that should be explored in further studies. Does the use of medical terminology result in more successful searches? Or does it result in documents whose language is inaccessible to health consumers? Are these terms used more often in familiar tasks? Although we found contradictory results in Section 6.6, the rare use of this type of terminology by health consumers opens doors in its exploration on relevance feedback techniques, assuming the language of the documents retrieved is still accessible to the user.

We found out that women, users that did previous searches on the topic and users that frequently use the Web to search for health and other types of information, use more terms per query. Are women more expansive in web search than men? The fact that users with greater familiarity express their information needs with longer queries agrees with some studies mentioned by Jansen and Pooch [13]. However, longer queries did not result in a larger situational relevance as described in Section 6.6.

Queries associated with psychiatry information needs have less terms than other specialties. Is it because it is harder to express psychological symptoms than physical ones? The overview and treatment types are also associated with shorter queries. We suppose it might be motivated by a desire of a larger recall in this exploratory kind of questions.

A clearer task is associated with longer queries. We think it is because users have a more clear idea of what they want and therefore think of more terms to describe the information need. Results also suggest a tendency to use longer and structured queries in more complex tasks.

Results report that younger and healthier users often classify documents as non-relevant. Is this type of users more strict in their criteria? Or does this happens because health searches are not so meaningful to this type of users and so they had less motivation to carefully evaluate the documents? We also found that male users judge relevance with higher scores.

Users with less years of web search experience tend to rate documents more often with 0 than with 1. Does this mean this type of users have less confidence in Web documents?

Users that usually conduct health searches on the Web tend to rate relevance higher. Yet, a frequent health searcher is more demanding than an occasional one, being associated with lower relevances. Interestingly, users find documents more relevant if they are using a familiar search engine. This suggests habit leads to trust.

We found out that users with previous searches on the topic tend to rate relevance lower. This result is in agreement with Saracevic [19]: "less subject expertise seems to lead to more lenient and relatively higher relevance ratings".

As expected, more difficult tasks have, in general, lower relevance scores. We also found that psychiatry has higher relevance when compared to other specialties. Has the Web more and better information on this topic? Is it because it is a topic easier to discuss online than in face to face conversations? About the question type, we found that the Prevention/Screening and the Prognosis/Outcome categories have the lowest relevance scores. The last result is not a surprise since it is hard to do a prognosis without a complete health profile.

Relevance is significantly higher when queries have advanced operators and use lay terms instead of technical ones. This last result contradicts a previous finding that says that the use of professional terminology is associated with a higher feeling of successful health searches (`hs_success`). With this result we see that situational and motivational relevance are not always in harmony. This is emphasized by another finding that says that users with a greater feeling of success have lower relevance scores. This jeopardizes evaluations done with the laboratory model and asks for evaluation models that incorporates other kinds of measures of success.

As expected, relevance decreases with the position of the document in the ranking. This finding agrees with the concept of ranking that is supposed to be ordered by relevance and with what Saracevic [19] reports: "information objects presented early have a higher probability of being inferred as relevant". We also found that pdf documents have higher relevance than html documents.

8. CONCLUSION AND FUTURE WORK

We have conducted a user study to analyze the influence of user and task context features on query formulation. Moreover, we analyzed the influence of the above features and also of query and document features on relevance judgments. We have reached findings that can foster new ideas to improve information retrieval and also ask for alternative measures of success.

Through the questionnaires we have asked users to evaluate their success rate in web search in health search and in the completion of the tasks in which they were involved. Some of our findings based on these variables were contradictory to the findings we have reached based on relevance judgments. This suggests traditional ways to evaluate IR systems can be improved through the incorporation of additional measures.

Our findings show that the use of advanced operators is directly connected with web search experience and that they lead to web and health search success. Similarly, the use of professional medical terminology is associated with familiarity with the topic and also leads to higher rates of successful health searches. Along with the rare use of professional terminology by health consumers, these findings can be used to detect expertise and adjust the IR process, applying specific query expansion techniques or adjusting the result sets.

Results have also raised hypothesis that should be tested in new studies, ideally focusing on a smaller set of variables to avoid interdependencies. A first hypothesis is that questions of the type Prognosis/Outcome need more user context to be successful. The other is that the Web is rich in psychiatric information and its anonymity attracts health searches on this topic.

Although english queries led to lower relevance scores, we believe the translation of terms to their english synonym might be a good strategy to improve the result set to users that understand English. We think the results of this study were affected by the low english literacy of the users. A future study could explore further the use of the English language in health searches by portuguese health consumers.

It would also 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. Also, other documents' characteristics like their language and their readability by lay people could be added to this analysis. An analysis similar to the one done in this study to situational relevance could also be done in terms of motivational relevance.

9 REFERENCES

[1] A. Aula. Query formulation in web information search. In *Proc. IADIS International Conference WWW/Internet*, volume I, pages 403–410, 2003.

[2] C. L. Barry. User-defined relevance criteria: an exploratory study. *J. Am. Soc. Inf. Sci.*, 45(3):149–159, April 1994.

[3] R. Bierig and A. Göker. Time, location and interest: an empirical and user-centred study. In *IIiX: Proceedings of the 1st international conference on Information interaction in context*, pages 79–87, New York, NY, USA, 2006. ACM.

[4] P. Borlund. The concept of relevance in IR. *J. Am. Soc. Inf. Sci. Technol.*, 54(10):913–925, May 2003.

[5] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.

[6] R. J. W. Cline and K. M. Haynes. Consumer health information seeking on the Internet: the state of the art. *Health Educ. Res.*, 16(6):671–692, December 2001.

[7] A. K. Dey and G. D. Abowd. Towards a Better Understanding of Context and Context-Awareness. In *CHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness*, 2000.

[8] P. Dourish. What we talk about when we talk about context. *Personal Ubiquitous Comput.*, 8(1):19–30, February 2004.

[9] S. Fox and S. Jones. The social life of health information. Technical report, Pew Internet & American Life Project, June 2009.

[10] D. J. Harper and D. Kelly. Contextual relevance feedback. In *IIiX: Proceedings of the 1st international conference on Information interaction in context*, pages 129–137, New York, NY, USA, 2006. ACM Press.

[11] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer, 1 edition, September 2005.

[12] P. Ingwersen, K. Jelin, and N. Belkin, editors. *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*, Royal School of Library and Information Science. Denmark., August 2005.

[13] B. J. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):235–246, February 2001.

[14] J. Lin and D. D. Fushman. Representation of information needs and the elements of context: A case study in the domain of clinical medicine. In *ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*, 2005.

[15] C. T. Lopes. Context features and their use in information retrieval. In *Third BCS-IRSG Symposium on Future Directions in Information Access*, September 2009.

[16] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[17] T. Saracevic. Relevance reconsidered. In *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*, pages 201–218, October 1996.

[18] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part II: nature and manifestations of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):1915–1933, November 2007.

[19] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part III: Behavior and effects of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):2126–2144, 2007.

[20] L. Schamber. A re-examination of relevance: toward a dynamic, situational definition*. *Information Processing & Management*, 26(6):755–776, 1990.

[21] L. Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology (ARIST)*, pages 3–48, 1994.