HAL open science

Analysis of biomedical and health queries: Lessons learned from TREC and CLEF evaluation benchmarks

Lynda Tamine, Cécile Chouquet, Thomas Palmer

To cite this version:

Lynda Tamine, Cécile Chouquet, Thomas Palmer. Analysis of biomedical and health queries: Lessons learned from TREC and CLEF evaluation benchmarks. Journal of American Society for Information Science and Technology (JASIST), 2015, 9 (1), pp.1+. hal-00940285

HAL Id: hal-00940285

https://hal.science/hal-00940285v1

Submitted on 20 May 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Manuscript JASIST–Submitted-12-2013

Analysis of biomedical and health queries: Lessons learned from TREC* and CLEF** evaluation benchmarks

Lynda Tamine · Cécile Chouquet · Thomas Palmer

the date of receipt and acceptance should be inserted later

Abstract A large body of research work examined, from both the query side and the user behaviour side, the characteristics of medical and health-related searches. One of the core issues in medical information retrieval is diversity of tasks that lead to diversity of categories of information needs and queries. From the evaluation perspective, another related challenging issue is the limited availability of appropriate test collections allowing the experimental validation of medically task oriented IR techniques and systems. Since 2003, medical standardized evaluation benchmarks such as TREC and CLEF provide to information retrieval research community various controlled medical tasks specifications with related document collections, queries, relevance assessment and specific metadata. The literature clearly reports a rapid increase in the use of these evaluation benchmarks. In this paper, we explore the peculiarities of TREC and CLEF medically oriented tasks and queries through the

* Text REtrieval Conference

** Cross Language Evaluation Forum

L. Tamine

Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, France, E-mail: tamine@irit.fr.

C. Chouquet

Université Paul Sabatier, Institut de Mathématiques de Toulouse, France, E-mail: cecile.chouquet@math.univ-toulouse.fr

T. Palmer

Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, France, E-mail: palmer@irit.fr.

analysis of the differences and the similarities between queries across tasks, with respect to length, specificity and clarity features and then study their effect on retrieval performance. More specifically, we developed an exploratory data analysis as well as a predictive data analysis using 11 TREC and CLEF medical test collections containing 374 queries and their corresponding document collections and expert relevance assessments, organized according to the involved medical tasks. Based on the outcome of our study, we show that, even for expert oriented queries, language specificity level varies significantly across tasks as well as search difficulty and that the most related predictive factors are linked to query length and query clarity. Additional findings highlight that query clarity factors are task dependent and that query terms specificity based on domain-specific terminology resources is not significantly linked to term rareness in the document collection. The lessons learned from our study could serve as starting points for the design of future task-based medical IR frameworks.

Keywords Information retrieval, query analysis, biomedical, health related queries, query difficulty, TREC, CLEF.

1 Introduction

The considerable amount of medical information available on different electronic resources has led to an increasing need of their access by users with health concerns (Zickuhr, 2010; Purcell, Rainie, Mitchell, Rosenstiel, & Olmstead, 2010). Furthermore, several studies have shown that the information obtained from medical and health-related searches can affect users' decisions about their self health or the health of people they care for (Fox & Jones, 2010; White & Horvitz, 2009). Given these findings, it is clear that information retrieval (IR) applications have much to gain from better understanding the user's health-related information needs hidden behind the queries with the light of context surrounding the search task in order to provide them with the most suited and relevant information. To facilitate research advances toward the achievement of this obejective, the IR research community provided, since 2003, standardized test collections within laboratory-oriented evaluation benchmarks, namely Text Retrieval Conference (TREC, www.trec.nist.gov) (Voorhees & Harman, 2005) and

Cross Language Evaluation Forum (CLEF, www.clef-initiative.eu). The medical retrieval tasks designed within these two benchmarks have provided a series of challenge evaluation and a platform for the presentation of their results. Roughly speaking, medical TREC and medical CLEF evaluation tracks provide 1) the description of medical tasks expressed by medical experts, 2) (bio)medical information needs, 3) a collection of documents being searched, and 4) a set of relevant documents per query, assessed by domain experts. Accordingly, the participants' challenge is to develop IR techniques and models which better fit the retrieval task outcomes assessed as relevant by medical experts involved in the evaluation benchmark design.

The principal contribution of this paper is the comparison between laboratory-evaluation query sets provided within TREC and CLEF benchmarks accross different medical tasks, according to a list of query features that characterize medical and health related searches, as reported through the literature. We achieve this by performing a statistical analysis using a set of predefined TREC and CLEF evaluation search tasks (gene retrieval, clinical case retrieval, diagnosis and search of links between an entity and a medical process) according to common features, namely query length, query clarity and query specificity and then evaluate their effect on query difficulty. In summary, outcomes drawn from our study may highlight more generally the medical tasks peculiarities and the predictive factors of their difficulty.

The remainder of this paper is structured as follows. In section 2, we give an overview of laboratory-style evaluation in IR. We present a review of medical and health-related searches in section 3. In section 4, we describe the key aspects of our research study; more specifically we detail the TREC and CLEF data tests, the main query features and the analyzed retrieval tasks. In section 5, we detail the statistical study results. In section 6, we discuss the study findings and highlight the study limitations. We summarize and conclude in section 7.

2 Background: Laboratory-style evaluation in IR

The laboratory model of IR evaluation relies on a batch evaluation mode. It aims at measuring, through a controlled environment, the effectiveness of retrieval models and algorithms, expressing their ability to identify topical relevant documents to users viewed as black boxes. The evaluation framework consists in a test collection containing a document collection, a set of topical and predefined queries and a set of relevance assessments identifying the documents that are topically relevant to each query. IR research area has experienced great advancements through the traditional laboratory model initiated by Cleverdon (Cleverdon, 1967) in the Cranfield project II. This model is widely used in IR evaluation campaigns such as TREC launched in 1991 by the National Institute of Standards and Technology (NIST) and the Defense Advance Research Project Agency (DARPA) in USA, INEX (INitiative for the Evaluation of XML retrieval) launched in 2002, CLEF launched in 2000. Laboratory-style evaluation contributed undoubtedly to new theoretical retrieval models on a large scale of information types, various IR tasks within different settings. The underlying Cranfield evaluation paradigm is as far convenient as experiments are undertaken in a laboratory environment where users are abstracted by sets of queries and static relevance judgments. The main advantage of such evaluation framework consists in the fact that it supports replicable experiments (Ingwersen & Jarvelin, 2005). More specifically, the algorithmic relevance is objective and measurable, leading to valuable comparisons, abstracting users and tasks and allowing to evaluate the goodness of the system. However, the laboratory evaluation model is moderately unsuited to the evaluation of systems involving users within search situations or tasks, assessing relevance with, for instance, interests, expertise and time in mind. More precisely, the main limitations of such model come from the basic assumptions used around the concept of relevance which are not realistic. Indeed, algorithmic relevance, addressed within this framework, is the minimal level of relevance being overwhelmed by other relevance levels such as cognitive, situational and affective relevance (Borlund, 2003). Furthermore, several studies validated the fact that relevance assessments are not generalizable across users even for the same query and the same level of relevance (Sormunen, 2002).

3 Studies on medical and health-related information needs and searches

Many studies in the literature have investigated different aspects of medical and health related information seeking and retrieval. These studies generally relied on empirical evaluations conducted with samples of users with an attempt to investigate the users' information need peculiarities (Lykke, Price, & Delcambre, 2012; Zhang & Fu, 2011; Spink et al., 2004), query difficulty (Lykke et al., 2012; W. Hersh et al., 2002; Boudin, Nie, & Dawes, 2012), user behaviour (Dogan, Muray, Neveol, & Lu, 2009; Ely et al., 1999), the effect of context on search (Freund, Toms, & Waterhouse, 2005; White, Dumais, & Teevan, 2008; Cartright, While, & Horvitz, 2011; Lykke et al., 2012), the search accuracy, the quality and the reliability of medical information (Pandolfini, 2002; Moturu, Liu, & Johnson, 2008), etc. Related findings provide insights into medical information search activity and suggest implications for the design of improved medical IR systems.

To put our contribution in context, we review below, state-of-the-art studies that specifically focused on the analysis of medical and health related information seeking process from the information need side as well as from the users' behaviour side.

3.1 Analysis of users' information needs

Table 1 provides background information on the main studies that investigated users' information needs in the context of medical or health-related IR tasks and highlights the key related findings. We look more particularly at query-related criteria commonly examined and/or formally defined and measured in the literature, described below:

- *Goal behind the query*: understanding the users' information needs has a critical importance for identifying effective IR strategies. To address this challenge, several studies attempted to explore the users' common goals in a health-related search session through questionnaires and semantic analysis of the search results. These studies demonstrate that users seek a very wide and increasing number of health-related information aspects such as: disease management, diagnosis, drug dosing, knowledge updating, advice, etc.

- *Query structure and vocabulary*: this feature provides the basic clues on which common IR models rely (Baeza-Yates & Ribeiro-Neto, 1999), based on the well known word based query-document matching where both (queries and documents) are considered as bags of words. Studies in health-related information search highlighted mainly that query vocabulary contains misspelling errors and abbreviations that trigger the problem of word matching gap. Moreover, these studies demonstrated that query words do not always match medical terminologies which lead to the problem of semantic gap.
- *Query length*: several studies in IR suggest that the query length is a helpful feature to understand the search goal and match relevant documents. Consistent with other studies which not target a specific IR domain, studies in health-related search reveal that queries are generally short, and most of them do not exceed 4 words either those addressed to medical IR systems or those addressed to general web search engines.
- *Query difficulty*: this feature is linked to the success or failure of the search process launched by the query. A large body of related research shows that several search strategies, performed at the query level, may leverage the query difficulty such as the mapping between the query vocabulary and the medical terminology entries, the use of semantic facets, the use of additional query terms, etc.

3.2 Analysis of users' search behaviour

We focus here on the users' behaviour during a health-related search session. Table 2 presents a synthetic review of the main related studies and the major underlying findings. More specifically, we have investigated them through the following criteria:

- *Used resources*: several information resources are exploited by users (experts vs. non-experts) to undertake a search: computerized resources, such as medical databases and open document collections namely the web, or non computerized ones through social relations (friends and parents) or print resources.
- *Search strategies*: studies highlight that there are no specific peculiarities characterizing users seeking for health-related information in comparison to other general searches. The main strategies consist in refining the query, browsing and searching, examining the top results.

Criterion	Key findings
Search goal behind the query	Search for general health, drug, dosing, cause of symptoms, disease management (search for rare diseases, updates on common diseases), (differential) diagnosis search, referral guidelines, continuing professional development, personified and opinion oriented goals (personalized healthy lifestyles such as diet, nutrition, sexual health), advice (Ely et al., 1999; Cullen, 2002; Spink et al., 2004; Ketchell, Lailani, Kauff, Barak, & Timberlake, 2005; L. Baxter, 2008; White & Horvitz, 2009; Dogan et al., 2009; Cartright et al., 2011; Davies, 2011; Zhang & Fu, 2011; Zhang, 2012; Herskovic, Tanaka, Hersh, & Bernstam, 2007); search for people with similar conditions in social platforms (Zhang & Fu, 2011), a wide variety of queries on a large variety of topics without dominant search terms or topics (Herskovic et al., 2007)
Query structure, vocabulary	There are common semantic category associations (Dogan et al., 2009); Terms do not always match standard vocabularies (Zhang & Fu, 2011; Keselman et al., 2008), inappropriate terms due to misspelling errors and abbreviations are common (Zhang & Fu, 2011; Boden, 2009), a query may contain two or three subqueries covering different facets of the information need (Zhang & Fu, 2011)
Query length	Queries are generally short: 1.5-4 terms (Maghrabi, Coreira, Westbrook, Gosling, & Vickland, 2005; Zhang, 2013), 1.5-2 terms (Lykke et al., 2012), 1.79 to 4 with an average of 2.81 (Zhang, 2013), median length is 3 (Herskovic et al., 2007); less than 3 terms in the case of web search engines (Spink et al., 2004)
Query difficulty	Use of few and core semantic query facets has a positive impact on the retrieval performance (Lykke et al., 2012); query expansion do not enhance the retrieval performance in all the cases (Mu & Lu, 2010; Trieschnigg, Kraaij, & Schuemie, 2006); searching with simple search terms is not effective (Berland, Elliot, & Morales, 2002); query term variability in relation to terminological concepts and semantic coverage of the query impact retrieval effectiveness (Boudin et al., 2012); combining textual features and visual features improves the effectiveness of multimodal queries (Radhouani, Kalpathy-Cramer, Bedrick, Bakke, & Hersh, 2009)

Table 1: A Synthetic overview of empirical studies on query characteristics in biomedical IR.

- *Search process pattern*: unlikely, this feature, related to the user cognitive "movement" during the search, is specific to health-related search tasks such as diagnosis. The key findings concern: 1) the evidence-directed reasoning pattern that attempts to link symptoms to disorders, 2) the hypothetico-deductive reasoning approach widely adopted by physicians involving two phases of analysis (evidence-directed phase and hypothesis directed phase) and 3) escalation related to the user's mental model that attempts to build a link between common symptoms and serious diseases vs. between serious symptoms and benign disorders.
- *Search difficulty*: unlike query difficulty, here the search difficulty is addressed from the user side. Several studies demonstrate that user-related factors such as users' class (student, nurse, etc), users' experience and users' task are relevant predictors of search difficulty.

4 Research design

In this section we first introduce the research objectives and then present the experimental data used to achieve our study, the query features modeled within the statistical analysis and the studied retrieval tasks.

4.1 Research objectives

The research review presented above highlights the large number of studies that contributed to gain a broad understanding of the research on health and medical-related information seeking. In terms of queries and related information needs, studies mainly relied on open evaluation designs employing either predefined or open queries addressed to standard (domain-oriented or general) IR systems. Most of the studies relied on naturalistic observations collected through questionnaires, or data analysis built on the user's search sessions generated data. Regarding query effectiveness, measures of outcomes relied on qualitative user-driven feedbacks or measurable performance (such as precision or nDCG) with care of the task simulation principles.

In terms of users and users' behaviour, a large body of research studied the impact of demographic variables and several factors, such as educational level and experience, on the search outcomes; some other studies examined

Criterion	Key findings
Used resources	Use of print (non-computerized) human resources (textbooks, journal papers) and interpersonal (parents, friends doctors), electronic resources mainly MEDLINE; electronic medical databases in increasing use by doctors, sometimes while the patient is waiting; use of several web resources for the same search task: search engines such as Google for health, health providers, Web 2.0 sources use of a combination of available medical databases improves the search effectiveness (Alper, Stevermer, White, & Ewigman, 2001; Cullen, 2002; Maghrabi et al., 2005; Andrews, Pearce, Ireson, & Love, 2005; Zhang, 2012)
Search strategies	Search frequently with keywords on MEDLINE, using of metadata filters (language, date of publication, etc), boolean operators and profiles to limit the search result length; query enhancement, query reformulation (Lykke et al., 2012; Zhang, 2013), use of semantic components for faceted search (Mu, Ryu, & Lu, 2011; Lykke et al., 2012); begin the search with general search engines (Spink et al., 2004); generally examine the top web pages of results (generally the first five) or the first page (Zhang, 2012; Tomes & Latter, 2007); search strategies evolve with the increasing of domain-knowledge (Wildemuth, 2004); sequencing knowledge starting from focused domain-specific resources (Bhavnani, 2001); searchers may look at and use the facets more when the health condition is perceived as more severe outcomes, transition between searching and browsing for complex tasks (Kules & Xie, 2011); medical students tend to view more result sets but fewer references, better able than nurse practitioner students to convert incorrect answers into correct ones (W. Hersh et al., 2002)
Search process pattern	Hypothesis-directed search (verifying hypothesis, narrowing search with a hypothesis, searching without hypothesis)(White & Horvitz, 2009); evidence-directed (building mental models on signs and symptoms with relation with disorders) (White & Horvitz, 2009); hypothetico-deductive reasoning running in two ordered phases: evidence-directed then hypothesis-directed (Eastin & Guinsler, 2006); escalation from common symptoms to serious illness Vs. rare symptoms to benign explanations (White & Horvitz, 2009); trial-and-error process (Tomes & Latter, 2007).
Search difficulty	Training and experience with a medical search engine lead to improved results; user type (student, nurse, etc) (W. Hersh et al., 2000, 2002; Pao et al., 1994) and user's task (W. Hersh et al., 2002; Inthiran, Alhashmi, & Ahmed, 2012; Zhang, 2013) are relevant predictors of search success.

Table 2: A Synthetic overview of empirical studies on user's behaviour in medical related search.

users' mental models when seeking for information for more focused tasks such as diagnosis.

These numerous studies obviously provide specific outcomes that can not be generalizable for all users in all information seeking situations. Different from previous work, we present, in this paper, a large-scale exploratory analysis of medical and health-related searches held specifically within TREC and CLEF evaluation campaigns. More precisely, the queries are expressed by experts and are drawn from a large set of data built upon a decade of research in IR. We focus particularly on the comparison between query features across various tasks being the context of the search (retrieve cohorts, retrieve patient cases, retrieve genes). We particularly studied three query features namely length, specificity and clarity, including each of them different facets related to different formal and operational definitions supported by different intuitions. The choice of these features is motivated by the research review presented above. Length is a relevant and common studied feature. Specificity gives insight on the query structure and vocabulary as well as the deepness degree of the expert's information need while clarity is intuitively linked to query difficulty (Steve, Zhou, & Croft, 2002). Moreover, exploiting the availability of the ground truth, we deepen our understanding on the impact of query features on the search effectiveness, regarding the involved tasks.

The specific research questions addressed in this study were as follows:

1. What are the differences vs. similarities between queries across tasks, according to length, specificity and clarity features?
2. What are the significant correlations between query features across tasks?
3. Which features result in low performance queries considering the various tasks?

To answer these questions, we carried out a statistical analysis based on 367 queries issued from eleven (11) TREC and CLEF evaluation data sets provided to the IR research community during a decade (from 2003 to 2012). We studied variables including queries, features, facets and search performance.

4.2 Experimental setup

Within the IR tasks analyzed in the paper, we only retain the following general and controlled setting:

- an ad hoc search task use case where the IR system aims to provide documents from the collection that are relevant to a hypothetical user information need expressed through a user-initiated query; in all the studied tracks, users are experts performing a simulated search in order to answer a simulated (not real) domain-oriented task: (a) medical professionals interacting with patients to suggest diagnosis or treatments, (b) genomic researchers seeking documents that describe how genes contribute to disease organisms;
- a natural language based query expression excluding metadata;
- the system output for a query is a list of ranked documents in a full text form. The document ranking is built based on the algorithmic relevance score considering the query under evaluation;
- relevance assessment, consisting in assigning relevant documents to each topic, is made by domain-experts. The overall relevance assessment made available through the data set is built according to the TREC pooling procedure (Harman, 1993).

Below, we give an overview of the evaluation campaign policies and then detail the method used for domain-concept recognition; afterwards, we describe the different query features analyzed in our study.

4.2.1 Data test collections

Text REtrieval Conference (TREC): TREC is a premier IR conference that specifies evaluation benchmarks with the objective of setting a baseline comparison across different research groups over the world. It has been organized by the U.S. National Institute of Standards and Technology (NIST, http://www.nist.gov/) (Voorhees & Harman, 2005) since 1992 and provides to participants different controlled tasks' specifications with related textual document collections, queries, relevance assessments and other metadata depending on the task being evaluated. Historically, in the area of medical and health-related IR, TREC has evolved since 2003 through the following tasks:

1. *TREC Genomics:* the TREC Genomics task was one of the largest and longest running challenge evaluation in biomedicine; it was launched in 2003 to answer first the challenge of managing and retrieving medical literature in order to identify potential interactions between genes, diseases and other biological entities (Radhouani et al., 2009). This task models the setting where a genomics researcher entering a new area expresses a query to a search engine managing a biomedical scientific literature, namely from MEDLINE collection. TREC genomics queries have evolved across years: gene names in 2003 (eg. "*arginine vasopressin*"), information needs expressed by genomic researchers entering a new area in 2005 (eg. "*provide information about the role of the gene PRNP in the disease Mad Cow Disease*") and question-answering in the biomedical domain in 2007 (eg. *"What is the role of gene gamma-aminobutyric acid receptors (GABABRs) in the process of inhibitory synaptic transmission?"*). Our study, presented in this paper, used the three different data collections including the three different forms of queries. The system should return the documents that are relevant for the considered query. While document collections were subsets of MEDLINE medical database abstracts in 2003 and 2005, the document collection included in 2006 full-text HTML documents from 49 journals that were electronically published via Highwire Press (Radhouani et al., 2009).
2. *TREC Filtering* (Robertson & Hull, 2000): this track aims at measuring the ability of an IR system to select relevant documents that fit persistent users' needs represented by profiles. It is worth to mention that we specifically used in our study the medical data set provided within this track for an ad hoc retrieval purpose, not a filtering one. More precisely, we made use of the OHSUMED test collection consisting of a set of 348,566 references from MEDLINE, the online medical database over a five-year journal (1987-1991) provided by W. Hersh (H. Hersh, Buckley, Leone, & Hickam, 1994). This collection is known as a large-scale standard collection for ad hoc medical IR (Stokes, Cavedon, & Zobel, 2009). We also used one of the subsets of the TREC-9 filtering track topics developed by Hersh and al. for their medical IR experiments (W. Hersh & Hickam, 1994). The ad hoc task simulates the use case assessing the use of MEDLINE by physicians in a clinical setting. The used topics include the patient information provided in a full text form (TI) and

the request description (AB), excluding Human-assigned MeSH terms (MH). An example of OHSUMED filtering query is "*adult respiratory distress syndrome*".

3. *TREC Medical:* this task ran during a couple of years (2011-2012) (voorhees & Hersh, 2012); the retrieval task consists in identifying cohorts in clinical studies for comparative effectiveness research. Queries specify particular disease/condition sets and particular treatment or intervention, expressed by physicians who are also students at Oregon Health and Sciences University in a natural language form; this document collection includes de-identified medical visit reports, made available for research use through the University of Pittsburgh Blulab NLP repository. Each report is associated with one or more medical visits. The system should return a list of visit reports that better fit the query specification. For example, a query might be "*find patients with gastroesophageal reflux disease who had an upper endoscopy*".

Cross Language Evaluation Forum (CLEF): The CLEF Initiative is a challenge evaluation for IR launched in 2000 with the aim of promoting multilingual and multimodal IR (Savoy, 2001). We used the following IR task:

1. *Medical image-CLEF task*: In 2004, CLEF organizers launched the Medical image-CLEF task (Muller et al., 2008) which focuses on the use of multimodal information (text and image) in the medical domain. Since 2009, a case-base retrieval task (sub-track), included in our study, has been running with the aim of promoting diagnosis retrieval; more precisely, the goal of the task is to retrieve cases including images that a physician would judge as relevant for differential diagnosis; roughly speaking, the main objective of the task is to provide to the clinician who expressed the query, a valuable aid to make a relevant diagnosis or treatment considering a difficult case. The queries were created from an existing medical case database including natural language based descriptions of medical cases including patient demographics, limited symptoms, test results and image studies. For example, a query might be: "*Female patient, 25 years old, with fatigue and a swallowing disorder (dysphagia worsening during a meal). The frontal chest X-ray shows opacity with clear contours in contact with the right heart border. Right hilar structures are visible through*

the mass. The lateral X-ray confirms the presence of a mass in the anterior mediastinum. On CT images, the mass has a relatively homogeneous tissue density". The system should return a list of articles from the literature that discusses similar cases. The used document collection is provided by the Radiological Society of North America (RNSA) and constitutes an important body of medical knowledge issued from peer-reviewed scientific literature (Muller et al., 2008).

Table 3 gives the statistics of the overall data (queries and documents) used in our study[1].

Year	**Queries (NB)**	**Documents (NB)**	**Relevant documents (NB)**
TREC Medical Track			
2011	35	95.701	1 765
2012	50	95.701	58 640
TREC Genomics			
2003	50	525. 938	566
2005	50	4. 591. 008	4 584
2007	36	162. 259	2 001
TREC Filtering			
1987	63	293. 856	3 205

Year	**Queries (NB)**	**Documents (NB)**	**Relevant documents (NB)**
CLEF Case base retrieval			
2009	5	5706	95
2010	14	77. 506	95
2011	10	55. 635	521
2012	26	74. 654	247
2013	35	74. 654	709

Table 3: Data set statistics

4.2.2 Domain knowledge extraction method

Biomedical text indexing techniques are faced with the well known problem of term identification (Krauthammer & Nenadic, 2004). In theoretical terminology, technical terms are used to define specialized *concepts* (Maynard,

[1] We notice here that we exclude from the study 7 queries from the 11 query collections because of the lack of associated relevance assessment.

2000). There are various definitions of a technical term as well as of a concept, but all agree that a term is essentially related to the linguistic realization of specialized concepts. Domain-based indexing relies on the use of semantic resources such as UMLS[2], MeSH[3], ICD-10[4], SNOMED[5], etc. In our study, we make use of the following means of text (either document or query) indexing:

- *The MeSH terminology*: our choice for this terminological resource relies on two main motivations: (a) previous work clearly shows that it is the most used general resource in the biomedical domain (Radhouani et al., 2009; Stokes et al., 2009), so it allows us to make the results issued from our study comparable to other ones issued from prior studies of the literature review; (b) since MeSH entails general biomedical terms, its coverage is wider compared to sub-domain specialized terminologies such as Gene Ontology[6]. Considering the use of different medical document data sets relying on different sub-domain knowledge (gene, medical, etc), this property allows us to limit the bias of domain characteristics on the study results.
- *Our IR based concept extraction method (Dinh & Tamine, 2011; Dinh, Tamine, & Boubekeur, 2013)*: more specifically, we use an IR-based approach for both MeSH concept categorization and document relevance estimation (Ruch, 2006). The key component of this method consists in representing the document semantic kernel as the top relevant concepts extracted by measuring the concept relevance for the document. Our basic assumption behind concept relevance is that a list of document words is more likely to map a concept that (1) the document and the concept both share a maximum number of words either among its preferred or non-preferred terms derived from all of its possible entries; (2) the words tend to appear in the same order so to cover the same meaning. Experiments carried out using different document collections in comparison with state-of-the-art concept extraction methods have shown its effectiveness either for mono-terminological (Dinh & Tamine, 2011) or multi-terminological indexing (Dinh et al., 2013).

[2] Unified Medical Language System
[3] Medical Subject Heading
[4] International Classification Disease
[5] Systematized Nomenclature of MEDicine
[6] www.geneontology.org/

4.2.3 Query features and facets

In our study, we consider an IR setting where an expert submits a query Q to a target collection of documents C. We consider three features to characterize the queries: length, specificity and clarity. Moreover, each feature is defined and formalized according to different intuitions or views, leading to the definition of different query facets. The notations and the measurements used for the definition of the feature facets are presented respectively in Table 4 and Table 5.

Notation	**Description**
std	standard deviation
$Words(Q)$	Set of query words
$Concepts(Q)$	Set of query concepts
n_w	Number of documents containing word w
N	Document collection size
$level(c)$	MeSH level of concept c
$sub(c)$	Set of subconcepts of concept c through MeSH "is-a" taxonomic relation
V	Document collection vocabulary
$P_{coll}(w)$	Relative frequency of word w in document collection *coll*
$D(Q)$	Set of documents that contain at least one query (Q) word
$P_{doc}(w\|d)$	Relative frequency of word w in document d
$P(d\|Q)$	Likelihood of a document (d) model generating the query Q (Song & Croft, 1999)

Table 4: Notations

- ***Query length.*** We consider the query length as a relevant attribute as investigated in previous work (Eysenbach & Kohler, 2000; Spink et al., 2004; Maghrabi et al., 2005; Dogan et al., 2009; Lykke et al., 2012). Furthermore, having in mind the fact that experts might make use of medical terminologies, we retain two facets regarding query components: (1) length as the number of stems or significant words (not

Variable	Facet definition	Measurement
	Length	
$LgW(Q)$	Query length in terms of words	$\|Words(Q)\|$
$LgC(Q)$	Query length in terms of concepts	$\|Concepts(Q)\|$
	Specificity	
$PSpe(Q)$	Posting specificity	$\frac{1}{LgW(Q)} \sum_{w \in words(Q)} - \log(\frac{n_w}{N})$
$ISpe(Q)$	Index specificity	$std_{w \in Words(Q)}(-\log(\frac{n_w}{N}))$
$HSpe(Q)$	Hierarchical specificity	$\frac{1}{LgC(Q)} \sum_{c_i \in Concepts(Q)} \frac{level(c_i)-1}{Maxlevel(MeSH)-1}$
	Clarity	
$WTCla(Q)$	Clarity through word distribution	$\sum_{w \in V} P(w\|Q) log_2 \frac{P(w\|Q)}{P_{coll}(w)}$ with $P(w\|Q) = \sum_{d \in D(Q)} P(w\|d)P(d\|Q)$ and $P(w\|d) = \lambda P_{doc}(w\|d) + (1-\lambda)P_{coll}(w)$
$CTCla(Q)$	Clarity through topic coverage	$\sum_{c \in Concepts(Q)} \|sub(c)\|$
$CCCla(Q)$	Clarity through concept coverage	$\frac{LgC(Q)}{LgW(Q)}$

Table 5: Query features and facets

including empty words), $LgW(Q)$, and (2) length as the number of terms referencing preferred entries of concepts issued from MeSH[7] terminology, $LgC(Q)$.

- ***Query specificity.*** Specificity is usually considered as a criterion for identifying index terms or descriptors (Jones, 1972). More specifically, we investigate three facets described below:
  1. *Posting specificity* $PSpe(Q)$: expresses the uniqueness of query words in the index collection; the basic assumption behind posting specificity is that the fewer documents involved by query words, the more specific the query topics are (Kim, 2006).

[7] MEdical Subject Headings

2. *Index specificity* $ISpe(Q)$: highlights the variation in information amount that query words carry (Pirkola & Jarvelin, 2001). Intuitively, we believe that the larger the index variation is, the less specific the query topics are.
3. *Hierarchical specificity* $HSpe(Q)$: it is based on the query words deepness of meaning defined in a reference terminology through the "is-a" taxonomic relation (Kim, 2006). The basic underlying assumption is that the more specific concepts are involved by the query words, the more specific the query topics are. Hierarchical specificity of a query is computed as the average normalized level of MeSH concepts that map the query words (Znaidi, Tamine, Chouquet, & Latiri, 2013).

- ***Query clarity.*** Broadly speaking, a clear query triggers a strong relevant meaning of the underlying topics, whereas an ambiguous query triggers a variety of topics meanings that do not correlate necessarily with each other. We usually distinguish between two classes of clarity scores: (a) pre-retrieval clarity scores are computed prior to retrieval, using query features such as length and query word frequency; (b) post-retrieval clarity scores are computed only after the query evaluation stage, using both query features and query retrieval result (document) features such as the distribution of query words in the result document list. We propose to compute the three following facets of the clarity feature:
  1. *Clarity through word distribution* $WTCla(Q)$: from a language modeling viewpoint, a query is clear if it returns a few different topics where a topic can be estimated by the distribution of query words over the result documents (Steve & Croft, 2002). Thus, the clarity score of a query is computed here at a post-retrieval stage, as the Kullback-Leiber divergence between the query language model and the collection language model.
  2. *Clarity through topic coverage* $CTCla(Q)$: a query is assumed to be as much clear as it covers a few general semantic levels of MeSH terminology (Znaidi et al., 2013). This score is computed at the pre-retrieval stage.

3. *Clarity through concept coverage* $CCCla(Q)$: a query is assumed to be as much clear as query words match concepts issued from MeSH terminology (Boudin et al., 2012). This score is also computed at the pre-retrieval stage.

4.3 Retrieval tasks

Guided by the objectives of the studied tracks on the one hand, and our research questions presented above on the other hand, we identified four (4) main different retrieval tasks. The difference between the tasks has been stated according to their difference with respect to a couple of fundamental dimensions: 1) task objective: the simulated medical IR scenario, 2) source of data : the information source acquired to achieve the task (scientific literature, patient medical records).

1. Task *T1 (NB. Queries= 100):* this task deals with gene retrieval and elicitation within the medical domain, supported by TREC Genomics 2003 and TREC Genomics 2005 evaluation campaigns. In 2003, topics were specifically gene names with the goal of finding relevant MEDLINE references that address biology or related protein products. In 2005, the task evolved towards the goal of answering genomic researchers' information needs put in the context of Generic Topic Templates (GTTs); these latter represented common medical backgrounds such as genes or diseases.
2. Task *T2 (NB. Queries= 85):* this task focuses on the retrieval of clinical cases, described through patient medical records, similar to the current one held by the query (case) at hand. This task is achieved through 2011 and 2012 TREC Medical campaigns.
3. Task *T3 (NB. Queries= 153):* this task simulates the situation where a physician seeks for relevant scientific articles that provide him with a fruitful assistance to achieve an accurate diagnosis/prognostic and/or to suggest a treatment considering the medical case at hand; this situation fits the well known Evidence Based Medicine (EBM) practice. This task is achieved through TREC Filtering as well as CLEF case base retrieval 2009-2013 evaluation campaigns.

4. Task *T4 (NB. Queries= 36):* this task concerns the search of typical relations between an entity and a medical process; it is based on the rationale that consumers of biomedical literature seek for relevant information and attempt to link between them as held within 2007 TREC Genomic evaluation campaign.

5 Results and discussion

We performed a statistical study in order to investigate the characteristics of medical and health related queries put in the context of predetermined tasks. The whole statistical analysis was carried out using the *SAS* (http://www.sas.com/) software (version 9.3), on 367 queries from the 374 queries initially contained in the collections. Indeed 7 queries were not used because of the lack of relevance assessment, which represented less than 2% of queries. This section details the used methodology and then presents and discusses the obtained results.

5.1 Query feature analysis

The objective of the first part of the study is to examine the characteristics of medical and health-related queries provided by TREC and CLEF IR tasks with an attempt to reveal possible correlations between them. The studied features are length, clarity and specificity as detailed above. The related variables are described using standard statistical indicators (mean, standard deviation, median, minimum and maximum) and their distributions were represented by box plots. The results are presented with respect to each task (T1-T4) and the corresponding means are compared across tasks using the non-parametric Kruskal-Wallis test[8] (Saporta, 2011). The correlations between the couples of features are evaluated using the Spearman correlation coefficient[9]

[8] The non-parametric Kruskal-Wallis test is performed in the case of k samples (k>2) coming from populations with similar characteristics. It is often used as an alternative to the analysis of variance when the normality assumption is not acceptable or when the numbers of observations are low.

[9] The non-parametric Spearman correlation coefficient is calculated between ranks of variable values rather than values themselves; it is used when the distribution of variables is not symmetric.

(denoted by *rho*) and are supplemented by a multivariate correlation analysis, namely the Principal Component Analysis (PCA[10]) (Saporta, 2011).

5.1.1 Do the query facets correspond to the different sides of the same feature?

We focus here on the analysis of the variations of the query facets belonging to the same feature. Our underlying objective is twofold: 1) examine the level of relatedness of the query facets (complementarity vs. overlap); 2) investigate whether the query facets relatedness is linked to the task specificities. In what follows, we report and discuss the obtained results grouped by feature.

- ***Query length.*** Figure 1.(a) and Figure 1.(b) show the query length variation per task, in terms of the number of words and the number of concepts respectively. We can see from Figure 1.(a) that, despite the fact that all the examined queries express experts' information needs, their lengths (noted by $LgW(Q)$) vary across tasks, with a range from 1 to 47; we notice particularly that queries belonging to task T3 are characterized by a large number of words (M[11] $= 16.8$, SD[12] $= 13.1$), unlike the other tasks for which queries are shorter: T1 ($M = 6.4$, $SD = 3.3$), T2 ($M = 6.7$, $SD = 2.7$) and T4 ($M = 6.6$, $SD = 1.9$). In addition, the analysis reveals that half the queries belonging to task T3 have more than 15 words while queries belonging to the other tasks contain a maximum of 16 words. This is explained by the length of the queries (despite the fact they were just a few) provided in CLEF case base retrieval track, created from the teaching file Casimage (Radhouani et al., 2009); they allow simulating the situation where a clinician diagnoses a difficult case described through a long description including patient demographics, limited symptoms, test results and image studies. The same overall trend is observed for the number of query concepts (noted by

[10] The PCA is a quantitative multivariate statistical method allowing to: (1) analyze the variability of data, and describe a maximum of variability, with new independent components that are linear combinations of the original variables, (2) explore the links between more than 2 variables and similarities between individuals, (3) synthesize information from a large number of variables.

[11] Mean

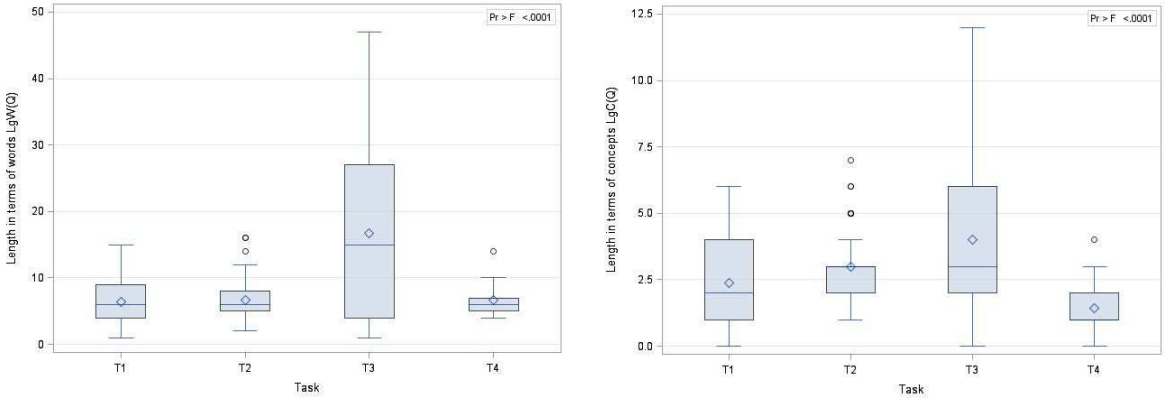
[12] Standard Deviation

$LgC(Q)$), ranging from 0 to 12 for all the queries across the different tasks, but does not exceed 7 for tasks T1 and T2, and 4 for task T4. We can see furthermore that all the queries belonging to task T2 map at least 1 concept, whereas for the other tasks, a low percentage of queries does not map any concept of MeSH: 2% for task T1, 3% for task T3 and in particular 14% for task T4. Considering the fact that concepts map mostly multi-words, two main reasons allow us to explain this finding: 1) experts use natural language in search and 2) given the fact that queries belonging to tasks T1 and T4 deal with genes, proteins and gene mutations, we expect that the coverage of general terminologies (such as MeSH) is however lower compared to gene specialized resources such as GeNE Ontology to point out specific terms; this gives rise to one of the limitations of this study that could be tackled by performing a multi-terminological concept recognition rather than a general terminology based indexation. The difference in the percentage of default-mapping is probably linked to the difference in the query lengths as observed in Figure 1.(a).

Moreover, the correlation study suggests that the two facets of length (in terms of number of words and number of concepts) are highly and positively correlated ($rho = 0.76$, p-value < 0.0001) indicating that a long query in terms of words is likely to map a large number of concepts whatever the task at hand. However, this trend is more pronounced in task T3: on average, queries belonging to task T3 contain in average 3.6 words per concept (95%CI[13]=[3.0-4.1]), vs. 2 (95%CI=[1.6-2.2]) for task T1 and 1.3 (95%CI=[0.6-1.9]) for tasks T2 and T4. This can be plausibly explained by the special use of a controlled language by physicians to describe the medical cases to be mapped to MEDLINE scientific resources, but we have to leverage the degree of word-concept mapping considering the large word lengths of queries belonging to task T3 as showed in Figure 1.(a).

- ***Query specificity.*** Figure 2.(a) shows the index specificity ($ISpe(Q)$) score variations across tasks. We can see that the $ISpe(Q)$ scores ranging from 0 to 0.35, are significantly different across tasks. It is worthwhile

[13] 95% Confidence Interval is a range of values that has a 95% chance of containing the true value of the estimated parameter; it is used to evaluate the accuracy of the estimate on a sample.



(a) Query length in terms of words ($LgW(Q)$)

(b) Query length in terms of concepts ($LgC(Q)$)

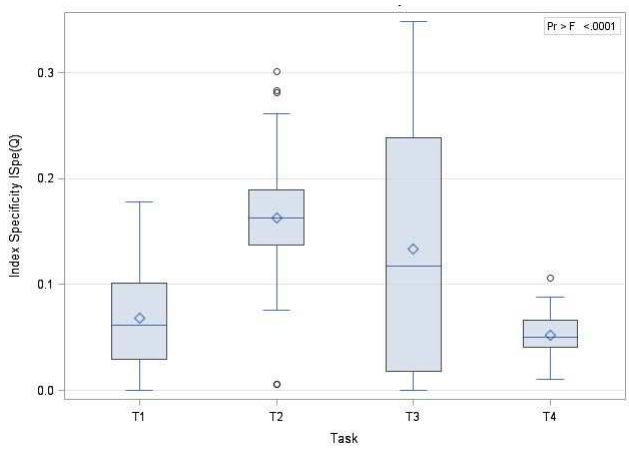
Fig. 1: Distributions of query length facets per task

to point out that queries belonging to tasks T1 and T4 have particularly low values of index specificity ($M = 0.07$, $SD = 0.05$ and $M = 0.05$, $SD = 0.02$ respectively), compared to queries belonging to tasks T2 ($M = 0.16$, $SD = 0.05$) and T3 ($M = 0.13$, $SD = 0.11$). This latter is also characterized by wider intervals of specificity values (0.35 maximum, Vs. 0.18 and 0.10 maximum for tasks T1 and T4 respectively). This can be explained by the nature of tasks T1 and T4 which involve information needs focused specifically on the biology of genes; within task T1, topics are basically gene names or their protein products, put moreover (in Genomics 2005 collection) in specific backgrounds leading to very focused queries; regarding task T4, the explanation relies, besides, on the nature of the question-answering task that seeks for high precision results and thus, leads to focused query topics. Regarding posting specificity ($PSpe(Q)$, in Figure 2.(b)), scores range from 0 to 10 with smaller differences across tasks: on average, 4.2 for task T1 ($SD = 1.4$), 3.2 for tasks T3 and T4 ($SD = 1.4$ and $SD = 0.5$ respectively), and 2.8 for task T2 ($SD = 0.9$). The lower difference across tasks may be explained by the fact that the facet is based on the degree of usage of "rare" words in the query rather on the variation of their rareness. The hierarchical specificity scores ($HSpe(Q)$, in Figure 2.(c)) range from 0 to 1 with wider and higher values for task T1 ($M = 0.37$, $SD = 0.17$) and homogeneous and lower values for task T2 ($M = 0.19$, $SD = 0.07$). This suggests that gene retrieval tasks make use of

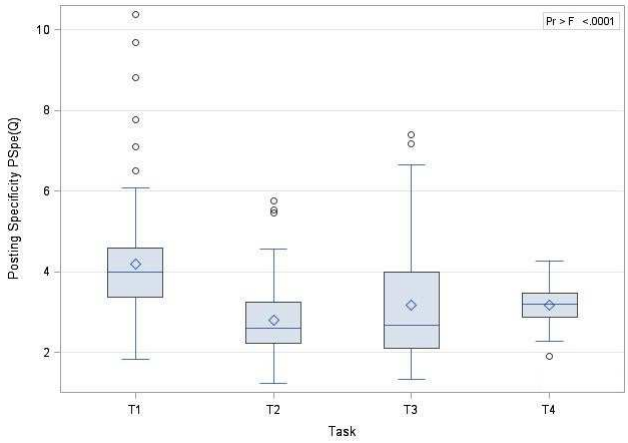
deeper (specific concepts) even if fewer concepts are recognized than the other tasks as revealed from the analysis of query length. The lower values observed specifically for task T2 suggest that the language used by physicians to identify similar cases is not at the deep level of MeSH and may be an attempt to make the mapping with other medical cases more successful.

Correlation analysis carried out on the three facets of specificity feature allows to highlight that posting specificity and index specificity are highly negatively correlated, regardless of the task ($rho = -0.82$, p-value< 0.0001). Thus, a query with high posting specificity has low values of index specificity, which is expected by the intrinsic definition of these facets: focused queries generally lead to the usage of words being at close levels of uniqueness, computed through their distribution (rareness) in the collection; inversely, broad query topics lead to the usage of words with different levels of uniqueness in the collection. In contrast, we observe that the hierarchical specificity, which constitutes the third facet, provides however another independent and complementary view of the specificity level of the queries. No significant correlation is observed between the hierarchical specificity and the two others facets: rho varies from -0.24 to 0.04 with index specificity (p-value> 0.05) and from -0.15 to 0.04 with posting specificity (p-value> 0.05). This suggests that specific terms, according to MeSH terminology, may be normally used to express expert medical and health information needs, and that they are normally distributed in the collection, ie. not particularly rare in the document collections.

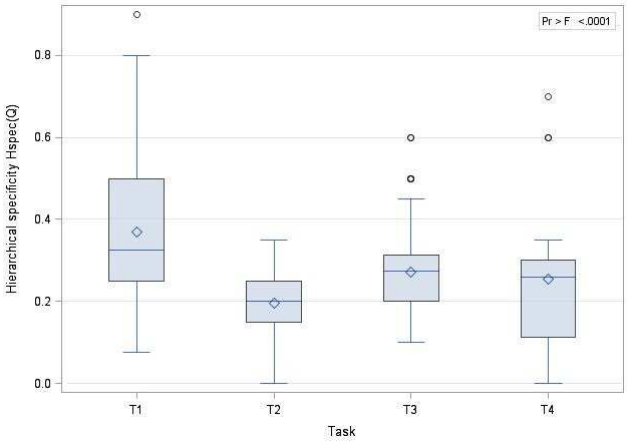
- ***Query clarity.*** The analysis of the three scores of query clarity gives quite different results as showed in Figure 3. More precisely, we can notice that the clarity through word distribution ($WTCla(Q)$, in Figure 3.(a)) contrasts lower values for tasks T2 ($M = 0.6$, $SD = 0.3$) and T3 ($M = 1.2$, $SD = 0.3$) with higher values of tasks T1 ($M = 1.7$, $SD = 0.6$) and T4 ($M = 2.0$, $SD = 0.4$). This suggests that health information needs are particularly less clear than the other natures of information needs such as gene retrieval. Indeed the distribution of the corresponding words in the collection is significantly different from their distribution



(a) Index specificity ($ISpe(Q)$)



(b) Posting specificity ($PSpe(Q)$)



(c) Hierarchical specificity ($Hspe(Q)$)

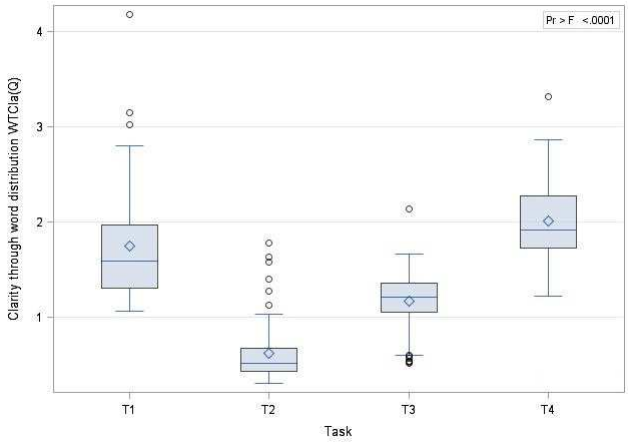
Fig. 2: Distribution of query specificity facets per task

in the documents being retrieved as an answer to the query. We can see however that the differences between tasks are less marked for clarity through topic coverage ($CTCla(Q)$, in Figure 3.(b)) with more heterogeneous values for task T3: half of the clarity scores for queries belonging to task T3 are below 3 ($M = 3.3$, $SD = 2.0$), against 75% of queries belonging to tasks T1 ($M = 2.1$, $SD = 1.1$) and T2 ($M = 2.5$, $SD = 1.0$) and 100% of the queries belonging to task T4 ($M = 1.4$, $SD = 0.9$). This can be explained by the fact that the use of MeSH concepts allows to normalize the description of the queries by reducing the effects of the

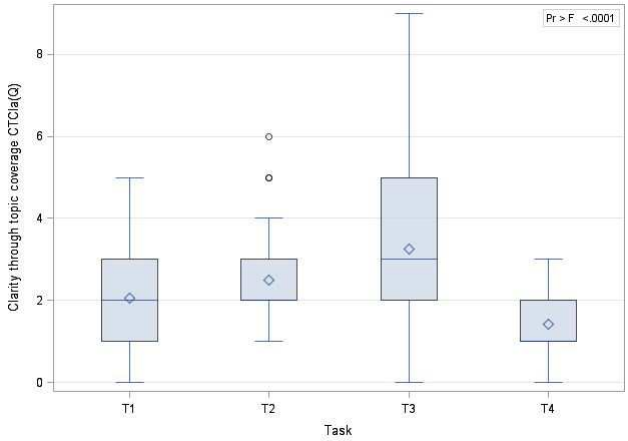
lexical differences that lead to the query word divergence as revealed by the precedent facet ($WTCla(Q)$). The heterogeneity of the scores may be caused by different factors: query length variation, default word-concept mapping and the nature of the task as well. The clarity through concept coverage ($CCCla(Q)$, in Figure 3.(c)) reveals larger differences between tasks, in particular between task T4 ($M = 0.21$, $SD = 0.12$) and task T2 ($M = 0.48$, $SD = 0.17$). For queries belonging to task T4, all the values of $CCCla(Q)$ are less than 0.4, against only 25% for queries belonging to task T2. Intuitively speaking, we can assume that health queries belonging to task T2 map MeSH entries more likely than medical queries belonging to task T4. Since the query length analysis reveals that they have close lengths, we guess that this observation is caused by the important difference variation in the mapping word-concept between these two groups of queries; thus, according to the facet view, T2 queries seem to be significantly clearer than T4 queries.

Correlation analysis between the query facets as well as the PCA showed different links between clarity scores according to the different tasks. Indeed, for queries belonging to tasks T1 and T3, a very strong significative negative correlation is observed between $WTCla(Q)$ and $CTCla(Q)$, regardless of $CCCla(Q)$ ($rho = -0.58$, p-value< 0.0001). Conversely, for tasks T2 and mainly T4, $CTCla(Q)$ and $CCCla(Q)$ facets are highly positively correlated, regardless of $WTCla(Q)$ facet ($rho = 0.43$ and $rho = 0.89$ respectively, p-value< 0.0001).

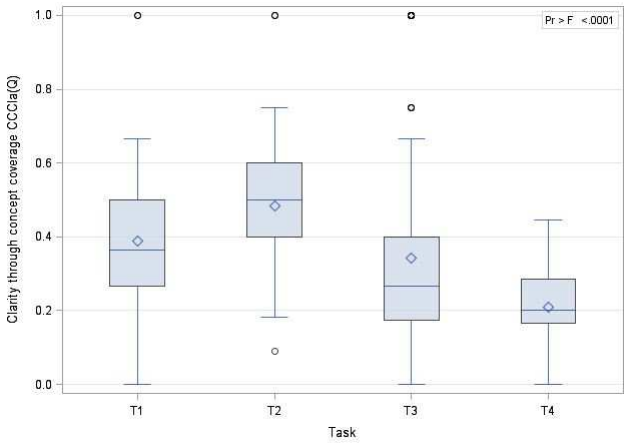
In summary, the study of the query characteristics shows a significant heterogeneity between the different tasks. Queries belonging to task T1 are characterized by higher values of posting specificity and hierarchical specificity. The queries belonging to task T2 have higher scores of index specificity and clarity through concept coverage, and lower scores of hierarchical specificity and clarity through word distribution. Task T3 has the largest number of queries (153) from six different document collections, which may explain the greater variability of the facets. This task is characterized by long queries in words and concepts, and the highest values of clarity through topic coverage. Conversely, task T4 contains only 36 queries from the same collection, which map with a very few concepts. These queries have the highest values of clarity through



(a) Clarity through word distribution ($WTCla(Q)$)



(b) Clarity through topic coverage ($CTCla(Q)$)



(c) Clarity through concept coverage ($CCCla(Q)$)

Fig. 3: Distributions of query clarity facets per task

word distribution and the lowest values of index specificity.

Our analysis also highlights pairs of facets of the same feature highly correlated: the lengths in terms of words and concepts are correlated positively, and index specificity and posting specificity are correlated negatively. Regarding clarity facets, clarity through topic coverage is correlated either positively with clarity through concept coverage or negatively with clarity through word distribution, depending on the considered task.

5.1.2 Correlation analysis of query features

Here, we focus on the study of the possible correlations between length, clarity and specificity by a PCA performed on the 8 facets according to each task, illustrated through Figure 4. We first observe that some links are highlighted whatever the task: it is the case for the positive correlation between lengths in terms of words and concepts, and the negative correlation between index specificity and posting specificity, which we have previously detected. We notice moreover that the length facets are highly and positively correlated with index specificity and negatively correlated with posting specificity. Furthermore, they are also positively correlated with the clarity facet through topic coverage. This indicates for example that longer queries have higher values for clarity through topic coverage and index specificity, but lower posting specificity values than shorter queries. The hierarchical specificity is highlighted independently of the other facets: correlation coefficients do not exceed 0.38, which reflects a lack of correlation between the hierarchical specificity and the 7 other facets. Finally, this analysis confirms that some correlations between facets are specific of some tasks, in particular for clarity. Indeed, the clarity through topic coverage appears negatively correlated with clarity through word distribution for tasks T1 and T3 (but not with clarity through concept coverage), whereas it is significantly positively related with the clarity through concept coverage for tasks T2 and T4 (but not with clarity through word distribution).

5.2 Impact of query features on the retrieval performance

The second part of the statistical analysis aims to establish possible relations between the query features and the retrieval performance by building a statistical model for identifying predictive features of query effectiveness.

5.2.1 Retrieval performance measurement

Evaluation of the query performance was based on traditional recall/precision measures. We ran the test queries on the document collections using two state-of-the art IR models namely the probabilistic Okapi BM25 model (Robertson & Jones, 1976) and the Hiemstra language model (Hiemstra, 2001), under Terrier search engine

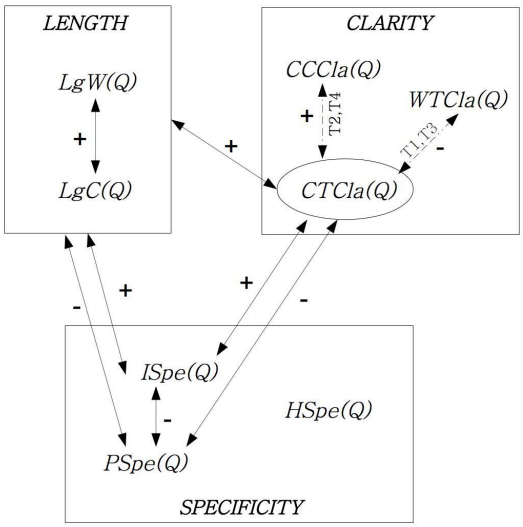


Fig. 4: Correlation analysis of query features

platform (http://terrier.org/). Based on the expert relevance assessments provided by each studied benchmark (*qrels*), we used the *Trec Eval* software (http://trec.nist.gov/trec_eval/) to compute the Mean Average Precision (MAP) performance metric, which is the mean of Average Precision (AP) computed over the set of test queries. This latter is computed for each query as the mean of Precision of each relevant item in a result list. Precision is computed as the proportion of relevant and retrieved documents out of the total number of retrieved documents. We notice that the retrieval performance computed for TREC Genomics 2007 (based on passage retrieval) was transformed first from the passage level to the document level, assuming that a document is relevant if it contains at least one relevant passage.

5.2.2 Difficulty threshold

In order to classify the queries either as *easy* or *difficult* according to their performance level, we estimated the probability density functions for the MAP scores of the queries (Duda & Hart, 1973; Steve et al., 2002).

The thresholds (called difficulty threshold), used to classify each query as *easy* or *difficult*, have been set up optimally based on the estimation of half (50th%) percentile of the Kernel density of the MAP scores; these thresholds can be interpreted as the median. Afterwards, if the MAP score is below the threshold, we classify a query as *difficult*, otherwise as *easy*. The difficulty threshold setting was run for each test collection belonging to each studied IR task (T1-T4). Two rankings (*easy* or *difficult*) are available, one by BM25 model, the other by the language model. As presented in Table 6, the concordance between those two rankings is excellent: in total, of the 367 studied queries, 365 have been classified in the same class by the two models. Only two queries (belonging to task T2) have not been classified identically by the two models: one in 2011 TREC medical track and another in the 2012 TREC medical track; for this coiple of queries, values of map scores are however very close to difficulty threshold. In addition, the Spearman correlation coefficient computed between the two MAP scores shows a significant strong correlation between them (rho=0.885 , p-value<0.0001). Based on these results, we used in the remainder of the analysis the classification based on BM25 model, knowing that our choice had no influence on the results presented later.

Number of classified queries	Using BM25 model	
Using Hiemstra Language model	*Easy*	*Difficult*
Easy	164	1
Difficult	1	201

Table 6: Comparison of difficulty rankings between BM25 model and Hiemstra language model

5.2.3 Descriptive analysis of query difficulty

Table 7 shows the statistics of MAP scores obtained within each document collection belonging to each IR task: the number of analyzed queries, the range of MAP values, the difficulty threshold and the number of easy queries

(and the corresponding percentage). We can see that all the indicators significantly vary depending on the task and moderately depending on the document collection. We can particularly see that the easier task is T2 with about 0.30 on median and quite half of the queries classified as easy. In contrast, tasks T1 (median=0.004), T3 (median=0.005) and T4 (median=0.05) seem to be less easy. We particularly highlight that the difficulty of TREC Genomics 2003 queries can be explained by the relevance pooling set up the year the IR task started (first year in 2003): documents were assessed as relevant if a Gene Reference into Function (GeneRIF) was available for the document. Giving the fact that the number of available GeneRIFs was small, the number of relevant documents is obviously underestimated (Radhouani et al., 2009). We can notice moreover that for CLEF case base IR tasks, the proportion of easy queries since 2012 is quite low (23% for 2012 and 29% for 2013), with a threshold fixed to 0, and a large number of queries having a null MAP score (77% and 71% of null values, respectively).

Task	**Collection**	**Number of analyzed queries**	**Range of MAP values**	**Difficulty threshold**	**Number (%) of easy queries**
T1	TREC Genomics 2003	50	$0 - 0.62$	0.0075	25 (50%)
	TREC Genomics 2005	49	$0 - 0.23$	0	22 (45%)
T2	TREC Medical track 2011	34	$0.016 - 0.87$	0.30	16 (50%)
	TREC Medical track 2012	47	$0.006 - 0.68$	0.32	24 (49%)
T3	CLEF Case base retrieval 2009	5	$0.008 - 0.67$	0.13	2 (40%)
	CLEF Case base retrieval 2010	14	$0.0009 - 0.64$	0.10	7 (50%)
	CLEF Case base retrieval 2011	9	$0 - 0.065$	0.008	4 (44%)
	CLEF Case base retrieval 2012	26	$0 - 0.25$	0	6 (23%)
	CLEF Case base retrieval 2013	34	$0 - 0.33$	0	10 (29%)
	OSHUMED	63	$0 - 0.40$	0.001	31 (49%)
T4	TREC Genomics 2007	36	$0.0003 - 0.53$	0.052	18 (50%)

Table 7: Analysis of query difficulty per IR collection

5.2.4 Predictive factors associated with query difficulty

Our objective here is to build a statistical model of the facets having a significant impact on the retrieval performance. In this model, the query performance is the response variable in two classes: "*easy*" vs. "*difficult*". We considered query features (length, specificity and clarity) as potential explanatory variables of the query performance. In this case, the appropriate method is the multivariate logistic regression, which allows to model the probability that a query is difficult according to query features (Christensen, 1997). To achieve this, we performed for each task (T1-T4) a first model based on the 8 features' facets, called *full model*. Significant predictive factors were selected using a step-by-step backward procedure from the full model: at each step, the factor which had the highest p-value was deleted from the model; this iterative procedure was re-run until all p-values associated to predictive factors were less than 0.05. Only factors which have a significant impact on query difficulty are included in the selected model. This means that factors not included in the selected model have not a significant effect on query difficulty. Furthermore, no interaction between predictors was significant. We use both Akaike Information Criterion (AIC)[14] and Bayesian Information Criterion (BIC)[15] to select the best model. (Saporta, 2011).

Table 8 presents, for each task, the results obtained from the best selected model: the set of significant predictive factors with the corresponding regression estimate and the associated p-value, the Odds Ratio (OR)

[14] The AIC score measures the fit quality of a statistical model, providing a means for model selection. It takes into account the goodness of fit of the model using the likelihood function and the model complexity (estimated through the number of parameters in the model); among a set of possible models, the preferred model is the one with the minimum AIC value

[15] The BIC criterion is close to the AIC one; they are complementary in selecting the best model. The only difference lies in how to measure the complexity of the model by a penalty term which is larger in BIC than in AIC

Significant predictive factors	**Regression Estimate**	**p-value**	**OR**	95% **CI**	% **Concordant Pairs**
Task T1 $(n = 99)$					
Length in terms of words $LgW(Q)$	0.64	< 0.005	1.90	$[1.28 - 2.82]$	72%
Length in terms of concepts $LgC(Q)$	-1.42	< 0.005	0.24	$[0.09 - 0.62]$	
Clarity through concept coverage $CCCla(Q)$ $(\times 10)$	0.43	0.05	1.54	$[0.99 - 2.37]$	
Task T2 $(n = 81)$					
Length in terms of words $LgW(Q)$	-0.24	< 0.05	0.79	$[0.63 - 0.99]$	78%
Hierarchical specificity $HSpe(Q)$ $(\times 10)$	-0.79	< 0.05	0.45	$[0.22 - 0.96]$	
Clarity through concept coverage $CCCla(Q)$ $(\times 10)$	-0.76	< 0.001	0.47	$[0.30 - 0.72]$	
Task T3 $(n = 151)$					
Length in terms of words $LgW(Q)$	0.04	< 0.01	1.04	$[1.01 - 1.07]$	61%
Task T4 $(n = 36)$					
Posting specificity $PSpe(Q)$ $(\times 10)$	-0.44	< 0.01	0.64	$[0.47 - 0.88]$	81%
Index specificity $ISpe(Q)$ $(\times 100)$	-0.59	< 0.05	0.55	$[0.31 - 0.99]$	

Table 8: Statistical model of query difficulty using a multivariate logistic regression per task

associated to each factor in order to quantify its effect on the query difficulty and its associated 95% Confidence Interval (95% CI), and the proportion of Concordant Pairs (CP), which allows to statistically evaluate the quality of regression. For further information on the interpretation of the effects of factors, an odds ratio greater than 1 means that the risk of difficult query is increased when the factor increases; conversely, when the odds ratio is less than 1, the higher the factor is, the easier the query will be. As showed in Table 8, for tasks T1, T2 and T4, the quality of the model is relatively good (around 80% of concordant pairs for T2 and T4, and 70% for T1). However, for task T3, the selected model contains only one predictor (namely the length in terms of words). The low percentage of concordant pairs (61%) indicates that query features do not allow to discriminate difficult queries from easy ones.

We can see that the predictive factors of query difficulty vary across tasks. For instance, for T1 task, three

factors have been selected as having a significant impact on the query difficulty (the length in terms of words and concepts, and the clarity through concept coverage), the other factors being pruned since they do not have significant effect on the query difficulty. For queries with the same number of concepts and the same clarity, we estimated that a term more in the query multiplied by 2 the risk of difficult query ($OR = 1.9$, p-value< 0.005). For queries with the same number of words and the same clarity, an additional concept divides by 4 the difficulty risk ($OR = 0.24$, p-value< 0.005). To a moderate extent, for queries of the same length (in words and concepts), the risk of difficult query is multiplied by 1.5 when the clarity (through concept coverage) increases by ten units ($OR = 1.5$, p-value$= 0.05$). In summary, for task T1, a query is even more difficult than it is long in terms of words, but short in terms of concepts with high clarity through concept coverage.

For task T2, the three selected predictive factors are correlated negatively with the query difficulty. The risk of difficult query increases with a lower number of words ($OR = 0.8$, p-value< 0.05), a lower hierarchical specificity $HSpec(Q)$ ($OR = 0.45$, p< 0.05) and mainly, a lower clarity through concept coverage $CCCla(Q)$ ($OR = 0.47$, p-value< 0.001). In summary, the shorter the query is in terms of words with low hierarchical specificity and clarity through concept coverage, the more difficult it is.

For task T4, the best selected model contains two predictive significant factors: posting specificity $PSpe(Q)$ and index specificity $ISpe(Q)$ for which higher values reduce the risk of difficult query ($OR = 0.64$, p-value< 0.01, and $OR = 0.55$, p-value< 0.05 respectively). In other words, a query of task T4 is even more difficult than it has low posting and index specificities.

6 Discussion

This study assessed the roles of length, specificity and clarity of TREC and CLEF queries across various medical tasks and their impact on search difficulty. We found various traits of similarities vs. differences across the test collections.

Regarding, query vocabulary, we mainly confirm two results: 1) terms do not always match standard vocabularies (Zhang & Fu, 2011; Keselman et al., 2008) through the analysis of both concept-based query length, and query clarity using concept coverage; 2) a query may entail sub queries covering different facets of the information need (Zhang & Fu, 2011) as showed through the analysis of query clarity through topic coverage, but also covers different levels of deepness regarding the terminological query word semantics, as revealed by the analysis of the hierarchical specificity.

The study also reveals that queries are quite long ranged from 1 to 47 with a bias towards queries particularly submitted within task T3 that focuses on typical clinical search queries. We argue that this trait is associated with the underlying IR predefined tasks and class of users submitting the queries (health experts). We believe that this query property is moderately realistic particularly in the case of health related queries addressing diagnosis or treatment search; users, mainly experts, submit longer search queries including natural language descriptions of symptoms, with the aim of retrieving more comprehensive relevant results (Luo, Tang, Yang, & Wei, 2007). Another major finding is the independency between hierarchical specificity and posting specificity as previously showed in (Kim, 2006). Thus, it would be interesting to explore search strategies that consider both word uniqueness and their semantic level in order to enhance the query structure. Besides, we showed in our study that query posting and query index specificities are highly positively correlated, suggesting that focused queries entail informative words. This may give evidence to perform search strategies aiming to reduce the sparseness of query topics or to perform local word disambiguation techniques. Another interesting result rising from the study is that clarity facets are task-dependent which results in avoiding their use as relevant indicators of the clarity of the underlying information needs, regardless of the task at hand.

The study findings give credence to previous work that argue that there is a strong relation between the task peculiarities and the search performance (W. Hersh et al., 2000; Inthiran et al., 2012; Zhang, 2013). Moreover, we found, that: 1) since word-based query length is not a relevant predictive factor of query difficulty in all the considered tasks, a long query does not lead to search effectiveness regardless of the tasks as previously showed (Mu & Lu, 2010); 2) with a moderate level of generalizability across tasks, we showed that the higher the degree

of word-terminological concept mapping, the lower the risk of query difficulty is; this result is consistent with a previous study that showed the high positive impact of the use of semantic components on the search results (Lykke et al., 2012). Moreover, our study highlights that index specificity reduces the risk of difficult queries, particularly in the case of focused medical questions as held in task T4. This result is at a moderate level, in line with previous work (He & Ounis, 2004), as authors showed the linear correlation between posting specificity and MAP scores but without however considering task contexts.

This study had some limitations. The experiments were limited to specific types of users, namely clinicians and medical professionals, expressing well defined queries as designed in up to date TREC and CLEF evaluation benchmarks. Moreover, we only used a general domain terminology, namely MeSH. Even if this latter allows to build a reasonable picture of the technical vocabulary of documents and queries, we believe that using multiple terminologies would allow us to deepen our understanding and interpretation of the result analysis. Another limitation was the low number of search tasks (T1-T4) and the difference in the number of topics that belong to each of them.

7 Conclusion

We have presented a study of medical and health-related searches centred on data sets provided by laboratory-based evaluation benchmarks well known in the IR community, namely TREC and CLEF. We analyzed four main search tasks and then examined the variation of query features (length, specificity and clarity) across tasks and their impact on search difficulty.

Our first research question focused on the analysis of query features and facets when searching for health information. Our findings suggest that, overall, the nature of the search task has a significant effect on the query characteristics. At the feature level, we observed that except for length, different facets of the same feature are not necessarily positively correlated, and that the direction of the correlation, if it exists, could depend on the

task at hand. The results suggest that: a) regardless of the task at hand, professionals which formulate queries on various topics use a vocabulary with a reasonable level of generality that limits the usage of rare (technical) words; however, we particularly observed that even if long, clinical queries do not entail much terminological concepts; b) clarity facets are task-dependent; more particularly, for gene retrieval and clinical diagnosis tasks (T1 and T3), it is likely that the more topically focused the queries expressed by the professionals are, the more different the specificity levels of the technical terms they use are; for clinical case retrieval and typical medical relation retrieval, it is likely that the more terminological concepts the professionals use to express their queries, the closer their MeSH specificity levels are.

Our second research question examined the nature of correlations between query features. We clearly observed that the hierarchical specificity measure, commonly used in medical IR through the usage of the semantic levels of domain terminologies, offers an independent level of query characterization rather than the commonly used query word inverse document frequency measure; this supports their use as complementary measures rather than commutable ones allowing to better gauge the query specificity according to both domain-knowledge and document collection topical scope. We also observed a positive correlation between length and index specificity and in contrast, a negative correlation between topical clarity and index specifity. This suggests that query expansion might involve a trade-off between query clarity and expanded-term specificity to ensure effectiveness improvement.

Our third research question asked about the features having a potential effect on medical or health-related query difficulty. The results reveal that the different tasks have different levels of difficulty, clinical queries being the most difficult ones. The results also suggest that, considering gene retrieval and medical case retrieval tasks, long queries are more difficult, with a level of difficulty depending on both the presence of terminological concepts and the overall query clarity, particularly for gene retrieval queries. Unlikely, searching for links between

medical entities depends heavily on the specificity of the involved entities descriptions.

The lessons learned from our study could serve as starting points for designing techniques that support task-based medical IR tools as well as designing experiments that leverage the importance of query features, namely length, clarity and specificity on the search success vs. failure. An interesting direction for future work would be to undertake a comparative analysis between expert vs. non-expert queries, using for instance the data issued from CLEF e-health track (Goeuriot et al., 2013). This would provide preliminary evidence of the diversity of users, variety of medical knowledge and variety of language skills involved witin medical IR tasks.

References

Alper, B., Stevermer, J., White, D., & Ewigman, B. (2001). Answering family physicians' clinical questions using electronic medical databases. *Journal of family practice*, *50*(11), 960-965.

Andrews, J. E., Pearce, K., Ireson, C., & Love, M. M. (2005). Information-seeking behaviors of practitioners in a primary care practice-based research network (pbrn). *Journal of the Medical Library Association (JMLA)*, *93*(2), 206-212.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Berland, G., Elliot, M., & Morales, L. (2002). Health information on the internet: accessibility, quality, and reliability in english and spanish. *Journal of the American Medical Association*, *285*.

Bhavnani, S. (2001). Important cognitive components of domain-specific knowledge. In *Proceedings of text rerieval conference trec* (pp. 571–578).

Boden, C. (2009). Overcoming the linguistic divide: a barrier to consumer health information. *Journal of the Canadian Health Libraries Association*, *30*, 75-80.

Borlund, P. (2003). The concept of relevance in ir. *Journal of the American Society for Information Science and Technology*, *54*(10), 913-925.

Boudin, F., Nie, J. Y., & Dawes, M. (2012). Using a medical thesaurus to predict query difficulty. In *Proceedings of the 34th international european conference on information retrieval (ecir)* (pp. 480–484).

Cartright, M., While, R., & Horvitz, E. (2011). Intentions and attention in exploratory health search. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (p. 65-74).

Christensen, R. (1997). *Log-linear models and logistic regression.* Springer.

Cleverdon, C. (1967). The cranfield test on index language devices. In *Aslib* (pp. 173–194).

Cullen, R. (2002). In search of evidence: Family practitioner's use of the internet for clinical information. *Journal of Americal Medical Library Association*, *90*(4), 370-379.

Davies, K. (2011). Physicians and their use of information: A survey comparison between the united states, canada and the united kingdom. *Journal of Americal Medical Library Association*, *99*(1), 88-90.

Dinh, D., & Tamine, L. (2011). Combining Global and Local Semantic Contexts for Improving Biomedical Information Retrieval. In *European conference on information retrieval (ecir)* (p. 375-386).

Dinh, D., Tamine, L., & Boubekeur, F. (2013). A study on factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies. *Artificial Intelligence in Medicine*, *57*(2), 155-167.

Dogan, R., Muray, G., Neveol, A., & Lu, Z. (2009). Understanding pubmed user search behavior through log analysis. *Database journal*, 1-19.

Duda, R. U., & Hart, P. E. (1973). *Pattern classification and scene analysis.* John Wiley and Sons.

Eastin, M., & Guinsler, N. (2006). Worried and wired: effects of health anxiety on information-seeking and health-care utilization behaviors. *Cybernetics and behaviour*, *94*(4), 494-498.

Ely, J., Osheroff, M. H., Bergus, G., levt, B. T., Chambliss, M., & Evans, E. (1999). Analysis of question asked by family doctors regarding patient care. *British medical journal*, *319*, 358-361.

Eysenbach, G., & Kohler, C. (2000). How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *Biomedical journal*, 573-577.

Fox, S., & Jones, S. (2010). *Generations 2010* (Tech. Rep.). The social life of health information Pew Internet & American Life Project. Retrieved from `http://www.pewinternet.org/`

Freund, L., Toms, E., & Waterhouse, J. (2005). Modeling the information behaviour of software engineers using a work task framework. In *Proceedings of the american society in information science and technology (assist)* (p. 1-10).

Goeuriot, L., Jones, G., Kelly, L., Leveling, J., Hanbury, A., Muller, H., et al. (2013). Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *Report on share/clef ehealth evaluation lab 2013.*

Harman, D. (1993). The second text retrieual conference (trec-2), gaithersburg, maryland, november 1993. In .

He, B., & Ounis, I. (2004). Inferring query performance using pre-retrieval predictors. In *In proc. symposium on string processing and information retrieval* (pp. 43–54). Springer Verlag.

Hersh, H., Buckley, C., Leone, T., & Hickam, D. (1994). Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international acm sigir conference on research and development in information retrieval* (pp. 192–201). New York, NY, USA: Springer-Verlag New York, Inc. Retrieved from `http://dl.acm.org/citation.cfm?id=188490.188557`

Hersh, W., Crabtree, M., Hickam, D., Sacherek, L., Friedman, C., Tidmarsh, P., et al. (2002). Factors associated with success in searching medline and applying evidence to answer clinical questions. *Journal of American Medical Information Association*, *9*(3).

Hersh, W., Crabtree, M., Hickam, D., Sacherek, L., Rose, L., & Friedman, C. (2000). Factors associated with successful answering of clinical questions using an information retrieval system. *Bull Medical Library Association*, *88*.

Hersh, W., & Hickam, D. (1994). Use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association*, *82*, 382-389.

Herskovic, J. R., Tanaka, L. Y., Hersh, W. R., & Bernstam, E. V. (2007). A day in the life of pubmed: Analysis of a typical day's query log. *Journal of Americal Medical Information Association*, *14*(2), 212-220.

Hiemstra, D. (2001). *Using language models for information retrieval*. Univ. Twente.

Ingwersen, P., & Jarvelin, K. (2005). *The turn: integration of information seeking and information retrieval in context*. Springer.

Inthiran, A., Alhashmi, S., & Ahmed, P. (2012). The affects of task difficulty on medical searches. In *Proceedings of the the world wide web conference (www)* (pp. 533–534).

Jones, S. (1972). A statistical interpretation of term specificity and its application to retrieval. *Journal of documentation*, *28*(1), 11-20.

Keselman, A., Smith, G., Divita, G., Kim, H., Brown, A., Leroy, G., et al. (2008). Consumer health concepts that do not fit that do not map to the umls: Where do they fit? *Journal of the American Medical Informatics Association*, *15*(15), 496-505.

Ketchell, D., Lailani, S., Kauff, D., Barak, G., & Timberlake, D. (2005). Prime answers: A practical interface for answering primary care questions. *Journal of Americal Medical Information Association*, *12*, 537-545.

Kim, G. (2006). Relation between index term specificity and relevance judgment. *Information Processing and management (IPM)*, *42*, 1218-1229.

Krauthammer, M., & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, *37*, 512–528.

Kules, B., & Xie, B. (2011). Older adults searching for health information in medline plus. In *Proceedings of the assist 2011* (p. 1Ŭ-10).

L. Baxter, E. H., N. Egbert. (2008). Everyday health communication experiences of college students. *Journal of American College Health*, *56*(4), 153-154.

Luo, G., Tang, C., Yang, H., & Wei, X. (2007). Medsearch: a specialized search engine for medical information. In *Proceedings of the* 16th *word wide web conference (www)* (p. 1175-1176).

Lykke, M., Price, S., & Delcambre, L. (2012). How doctors search: A study of query behaviour and the impact on search results. *Information Processing Management*, *48*(6), 1151–1170.

Maghrabi, F., Coreira, E., Westbrook, J., Gosling, A., & Vickland, V. (2005). General practitioners'use of online evidence during consultations. *Journal of Americal Medical Library Association*, *74*, 1-12.

Maynard, D. (2000). *Term Recognition using Combined Knowledge Sources.* Unpublished doctoral dissertation, Manchester Metropolitan University, UK.

Moturu, S., Liu, H., & Johnson, W. G. (2008). Trust evaluation in health information on the world wide web. In *Engineering in medicine and biology society, 2008. embs 2008. 30th annual international conference of the ieee* (p. 1525-1528).

Mu, X., & Lu, K. (2010). Towards effective genomic information retrieval: the impact of query complexity and expansion strategies. *Journal of Americal College Health*, *54*(8), 194-208.

Mu, X., Ryu, H., & Lu, K. (2011). Supporting effective health and biomedical information retrieval and navigation: A novel facet view interface evaluation. *Journal of Biomedical Informatics*, *44*(4), 576-586.

Muller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T., Clough, P., et al. (2008). Overview of the image clef 2007 medical retrieval and annotations tasks. In *Clef 2007 proceedings* (p. 473-491).

Pandolfini, C. (2002). Follow up of quality of public oriented health information on the world wide web: systematic re-evaluation. *British Medical Journal*, *324*, 582–3.

Pao, M., Grefsheim, S., Barclay, M., Wooliscroft, J., Shipman, B., & McQuillan, M. (1994). Effect of search experience on sustained medline usage by students. *Academic Medicine*.

Pirkola, A., & Jarvelin, K. (2001). Employing the resolution power of search keys. *Journal of the American Society for Information Science and Technology (JASIST)*, *52*(7), 575-583.

Purcell, K., Rainie, L., Mitchell, A., Rosenstiel, T., & Olmstead, K. (2010). *Understanding the participatory news consumer* (Tech. Rep.). Pew Internet & American Life Project.

Radhouani, S., Kalpathy-Cramer, J., Bedrick, S., Bakke, B., & Hersh, W. R. (2009). Using media fusion and domain dimensions to improve precision in medical image retrieval. In *Proceedings cross language evaluation framework (clef)* (pp. 223–230).

Robertson, S., & Hull, D. (2000). The tre-9 filtering track final report. In *Proceedings of trec 2000.*

Robertson, S., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society on Informtion Science and Technology (JASIST)*, *27*(3).

Ruch, P. (2006, March 15). Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, *22*(6), 658–664. Retrieved from `http://dx.doi.org/10.1093/bioinformatics/bti783`

Saporta, G. (2011). *Probabilités, analyse des données et statistique.* Editions Technip.

Savoy, J. (2001). Report on clef-2001 experiments. In *Report on the clef conference 2001* (p. 27-43).

Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the 22nd international acm sigir conference on research and development in information retrieval (sigir)* (p. 279-280).

Sormunen, E. (2002). Liberal relevance criteria of trec : counting on negligible documents? In *Proceedings of the* 25th *annual international acm sigir conference on research and development in information retrieval* (pp. 324–330). New York: ACM.

Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D. P., & Ozmutlu, S. (2004). A study of medical and health queries to web search engines. *Health information and libraries journal*, *21*, 44-51.

Steve, C. R., & Croft, W. (2002). Quantifying query ambiguity. In *Proceedings of the second international conference on human language technology research* (pp. 104–109). San Francisco, CA, USA. Retrieved from `http://dl.acm.org/citation.cfm?id=1289189.1289266`

Steve, C. R., Zhou, Y., & Croft, W. (2002). Predicting query performance. In *35th international acm sigir conference on research and development in information retrieval* (pp. 299–306).

Stokes, N., Cavedon, Y., & Zobel, J. (2009). Exploring criteria for succesful query expansion in the genomic domain. *Information retrieval*, *12*, 17-50.

Tomes, E., & Latter, C. (2007). How consumers search for health information. *Health informatics journal*, *13*(3), 223-235.

Trieschnigg, D., Kraaij, W., & Schuemie, M. (2006). Concept based document retrieval for genomics literature. In *Fifteenth text retrieval conference, trec 2006.*

Voorhees, E., & Harman, D. (2005). *Trec: Experiment and evaluation in information retrieval.* MIT Press.

voorhees, E., & Hersh, W. (2012). Overview of the trec 2012 medical records track. In *Proceedings of the trec 2012.*

White, R., Dumais, S., & Teevan, J. (2008). How medical expertise influences web search interaction. In *Proceedings of the 30th international acm sigir conference on research and development in information retrieval* (p. 791-792).

White, R., & Horvitz, E. (2009). Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems*, *27*(4), 1-37.

Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246-258.

Zhang, Y. (2012). Consumer health information searching process in real life settings. In *Proceedings of the american science in information technology (assist)* (p. 1-10).

Zhang, Y. (2013). Searching for specific health-related information in medlineplus: behavioral patterns and user experience. *Journal of the American Society for Information Science and Technology (JASIST).*

Zhang, Y., & Fu, W. T. (2011). Designing consumer health information systems: What do user-generated question tell us? In *Proceedings of the 6th international conference on foundations of augmented*

cognition: directing the future of adaptive systems (pp. 536–545).

Zickuhr, K. (2010). *Generations 2010* (Tech. Rep.). Pew Internet & American Life Project. Retrieved from `http://www.pewinternet.org/~/media//Files/Reports/2010/PIP_Generations_andTech10.pdf`

Znaidi, E., Tamine, L., Chouquet, C., & Latiri, C. (2013). Characterizing health-related information needs of domain experts. In *Proceedings of the artificial intelligence in medicine conference (aime)* (p. 48-57).