Context-Based Health Information Retrieval

Carla Alexandra Teixeira Lopes
Department of Informatics Engineering
Faculdade de Engenharia da Universidade do Porto
carla.lopes@fe.up.pt

ABSTRACT

Health Information Retrieval (HIR) applies the concepts and techniques of Information Retrieval (IR) to the actors, behaviors and information of the healthcare domain. Contextual Information Retrieval (CIR) is a subarea of IR that seeks to incorporate contextual features in the retrieval process towards its improvement. These areas have been gaining interest from the research community but it's unusual to find research crossing the two domains. This research will study the context of HIR, identify elements of context potentially significant to HIR and propose novel ways to improve HIR performance through the exploration of contextual features. This paper presents the motivation behind this work, a description of the proposed research, a literature review relevant to the research problem and the methodology to fulfill it.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Contextual information retrieval, health information retrieval, user context

1. INTRODUCTION

Typically, Information Retrieval (IR) systems support their decisions solely on the query and document collection. Several implicit factors about the user and the search context (e.g. time, location, task, expertise, interaction) are ignored and could be considered to optimize IR performance.

In domains like health, the search context is extremely rich. Here, the search process usually occurs in well-defined scenarios (e.g. "treatment", "diagnosis") whose characterization may be used to improve IR systems. These systems are largely used by both domain professionals and non-experts.

This research intends to study what elements of context are potentially significant to Health Information Retrieval (HIR) activities and to propose novel ways to improve HIR performance with the exploration of contextual features.

This document is organized as follows. It starts presenting the motivation behind this work in Section 2 and a statement of the research problem along with its aims and thesis in Section 3. It then includes an outline of the precedents for the proposed work with a literature review relevant to the research problem in Section 4 and a description of the proposed research methodology in Section 5. Finally, an enumeration of the research issues to be discussed at the Doctoral Consortium is presented in Section 6.

SIGIR '09 Boston, USA

2. MOTIVATION

HIR is the application of IR concepts and techniques to the healthcare domain. Just like the broader area of IR, HIR has largely evolved in the last few years. The increasing availability of information in a digital format [16] is having impact on the health domain where access to information has become easier to professionals and consumers (patients, their families and friends). Their greater demands is triggering research intiatives in HIR and the creation of specialized services and products (an extensive list thereof has been compiled by Lopes [15]). These events and the growing tendency on health information consumption have been forecasted and are expected to continue happening in the near future [10].

In parallel, *context* is gaining attention in the field of Information Retrieval (see Section 4). All information activities take place within a context that affects the way people access information, interact with a retrieval system, evaluate and make decisions about the documents retrieved [9, 12].

According to Dourish [6], *context* may be defined in two perspectives: as a representational problem or as an interactional problem. In the first perspective, it is viewed as a form of information that is delineable, stable and independent of activity. It is viewed as a set of implicit attributes that describe the user and the environment in which information activities occur. This is the perspective most commonly adopted in research papers that want to capture and represent a stable context [9]. The second perspective sees context as arising from the activity, from which it can't be separated.

Goker et al. [8] present a context model as part of the

AmbiSense research project. In this model, context elements are divided into five main categories: task (e.g.: what the user is doing, his goals, tasks, activities), social (e.g.: social aspects of the user, such as information about friends and family, his role), personal (e.g.: mental and physical information about the user such as mood, expertise, disabilities), spatio-temporal (e.g.: attributes like time and location) and environmental (e.g.: users surroundings like things, light, people and information accessed by the user).

Several authors agree that context, often ignored, might be used to improve the retrieval process [12, 3]. A contextualised IR could allow IR systems to learn and predict what information a searcher needs, learn how and when information should be displayed, present how information relates to other information that has been seen and how it relates to other tasks the user was engaged in and decide who else should get the new information.

According to Lin et al. [13], "the domain of clinical medicine is very well-suited for experiments in building richer models of the information seeking process". In fact, it's not difficult to foresee contextual features in this domain that could enrich HIR models. In the health domain, the search process usually occurs in well-defined scenarios (e.g.: "treatment", "diagnosis") [14] and context may be extremely rich. Similarly to any visit to the doctor, where the patient doesn't just say "itch", but explains the context of the "itch" to the doctor, context is relevant to HIR. Other examples of contextual features that can be used are: searcher's personal health record, clinical case in hands, searcher's expertise in the health domain.

Hersh, one of the major figures in HIR, pointed in a 2007's presentation that "finding the right information for the right task" is one of the grand challenges for HIR. Together with the relevance of context in the health field [19] and the spread of HIR activities, this stimulates this research.

3. PROPOSED RESEARCH

This research thesis may be defined as - "Context can be used to improve Health Information Retrieval". This research seeks to study how contextual features surrounding Health Information Seeking and Retrieval can affect the use of HIR systems and to apply these features in the improvement of these systems. More specifically, it will focus on:

1. The identification of potentially significant elements of context to HIR (e.g. task features, personal features, interaction features, social features),
2. Ways to capture these context features,
3. The improvement of HIR with IR strategies that use the identified contextual elements,
4. IR prototypes, developed on top of standard retrieval models, to test and evaluate the defined strategies.

The large quantity of contextual data that may be associated with an HIR system make us believe that context can be useful to deal with some of its known problems and to improve it. For example, it can be useful to increase the relevance of the returned documents helping to manage the vast amount of available health information. It may also be useful to deal with the problem of complicated medical terms that are commonly misspelled or incomplete in search engines' queries [19] with the use of task and personal context features. It's also possible to imagine ways to help health information quality assessment - another commonly mentioned problem [5] - through the use of social context features.

This research will be mainly dedicated to text content instead of other types of media content like images, videos and sounds. It will also be focused on some specific phases of the retrieval process. Figure 1 presents the generic architecture of the retrieval process, it is possible to distinguish the stages, by the boxes with grey filling, where this research will focus. The dotted lines represent search contextual features that feeds, implicitly or explicitly, the specified retrieval phases.

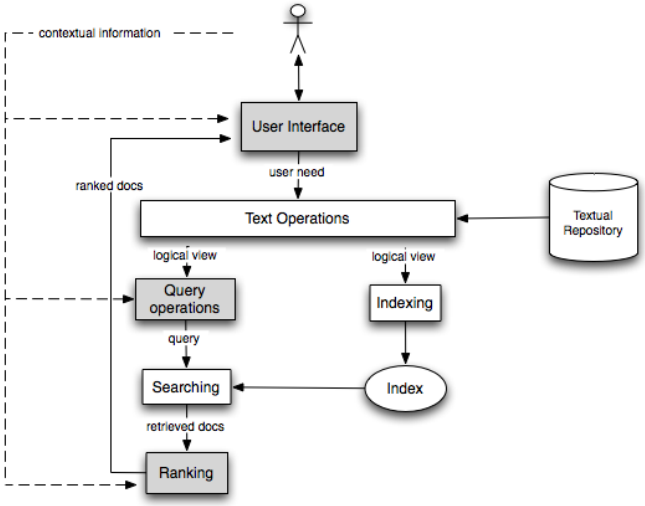

Figure 1: The Retrieval Process [adapted from [1]]

These retrieval phases were chosen because they have a great potential of improvement through the use of contextual information and are associated with fewer barriers than other phases like the indexing one. In fact, the required infrastructure and financial investment are not the same in all phases - to store a repository and implement an indexing process is expected to be more demanding in infrastructures and more dependent on specific datasets.

4. BACKGROUND AND RELATED WORK

In the field of Information Retrieval, the interest to adapt the search process towards the user needs and context has been growing [3]. Contextual IR, Adaptive IR and Interactive IR are names usually given to this research area. Several journals and conferences have given attention to this topic in the last few years. The Information Processing and Management journal has dedicated two special issues to this topic, one in 2002 and another in 2008. The Information Retrieval journal has also dedicated a special issue in 2007 to Contextual Information Retrieval Systems.

In 2004 and 2005 two workshops entitled Information Retrieval in Context (IRiX) were held in association with the ACM SIGIR Conference. In 2006, a new set of conferences entitled Information Interaction in Context (IIiX) has begun as a spin-off. Another workshop about Adaptive IR took

place in SPIRE'06. In 2007, a workshop entitled *Context-Based Information Retrieval* was held in the Sixth International and Interdisciplinary Conference on Modelling and Using Context. In 2008 occurred the second edition of IIiX, a conference called Information Seeking in Context (ISIC) and a workshop at Informatik 2008 entitled *Use In Context: Between adaptivity and adaptation in context-based interactions*. In 2009, a workshop at the 31th European Conference on Information Retrieval will be dedicated to *Contextual Information Access, Seeking and Retrieval Evaluation*.

All these journals and conferences provide an overview of the relation between context and information retrieval, present case studies, propose evaluation and research methodologies, offer new ways for modelling context and provide frameworks for doing research in this area.

The interest in HIR has also been growing [15]. Recent research works propose new ways to adjust queries to user needs, to index health data, to present search results, to rank results (e.g. based on readability, content) and to analyze information needs and behaviors [15]. There are also specific HIR research works that focus on the exploration of contextual features. From these, the ones that focus or may have impact at the retrieval process's phases included in this research (Interface, Query Operations and Ranking) will be presented next.

Maviglia et al. [18] conduct a study to determine the impact of a medication infobutton in health professionals' questions and in their decisions. This button, called KnowledgeLink, is embedded within other applications and returns patient-specific and context-specific links from medications that appear in patient's health records. They conclude that, although used infrequently, this button satisfies the clinician's information needs and has impact on decision making.

Silva et al. [19] use data from Personal Health Records (PHR) as a context to deliver health information search results adapted to the user health conditions. Their system uses Google to gather the user's query results and reorders them based on the occurrences of context keywords (gathered from the PHR) on these documents. The results' presentation is another component of the system: each result has a contextual line (with the terms of the PHR found in the document), document's cached version with highlighted context keywords is presented and some filtering options are presented in the results page. Results show that adding context to the retrieval process is useful when users are learning about a topic or when information is ambiguous and not so useful when users look for answers to specific questions.

Liu et al. [14] describe a query expansion method that has into account the scenario/task context (e.g.: "treatment", "diagnosis") in retrieving medical free text. This expansion method uses the Unified Medical Language System (UMLS) knowledge source to add terms relevant to the query's scenario. When compared with a traditional method that expands terms statistically correlated but not necessarily scenario specific, this method showed improvements in precision-recall.

Lin et al. [13] identify and describe five context elements that improve the representation of user's needs in queries: the work task model, the search task model, the process model, the problem structure and the domain model. The authors mention the use these elements in the building of a more refined model of the information seeking process.

Martins et al. [17] propose a context-aware information retrieval process composed of a semantically built index and a relevance feedback approach. The semantic index is built using natural language tools to analyze each document and link their terms to concepts of the UMLS ontology. The relevance feedback approach is based on implicit evidences provided by contextual information and explicit evidences provided by the user interaction while browsing the results.

White et al. [20] study the effect of domain expertise on web search behavior in four different domains (one of them is medicine). They characterize the nature of the queries, search sessions, visited web sites, and search success for users identified as experts and non-experts within these domains. With this analysis they build a model to predict expertise and show how this information may be used to improve IR systems (e.g. present better results, query suggestions, help non-experts gain expertise).

Yan et al. [21] deal with the problem of different readability levels in health documents and its adjustment to users with different domain knowledge. To allow different rankings for health-experts and non-health experts they propose a model of text readability. The model takes into account two factors: how the domain-specific concepts affect the documents readability and the document's textual genres. They propose and apply (in the context of HIR) three readability formulas. When compared to four traditional readability measures, the proposed formulas showed improvements.

5. PROPOSED METHODOLOGY

This research is expected to be multidisciplinary. Besides the IR field, the competencies and techniques from areas like Knowledge Management, Data Mining, Machine Learning, Natural Language Processing, Human Computer Interaction and Information Seeking Behavior, will be needed at different stages.

An important aspect to the success of this research is the access to relevant datasets. Contacts have already been made towards an easy access to several national hospitals. This will facilitate the analysis of health professionals' information behavior, the acquisition of relevant datasets (e.g.: query and access logs, surveys' answers) and will allow the evaluation of prototypes. The diversity of hospitals will allow the creation of a larger and richer sample of cases and, therefore, a better characterization of health information seeking behavior.

Other datasets have already been obtained. Some of them are: web access logs from the computers at an hospital's library, set of access logs from a well-known portuguese search engine, set of queries classified in categories (one of them is health) and the three knowledge sources of the Unified Medical Language System. These datasets will be useful to characterize health information seeking behavior and, directly or indirectly, will be useful to the three phases of retrieval process where this research will act.

The defined methodology includes several steps that will be briefly described next.

5.1 Information Seeking Behavior Studies

The characterization of Information Seeking Behavior is usually a prerequisite to the creation of an IR system [3]. Therefore, an Health Information Seeking Behavior study is needed to find the context attributes that matter for HIR applications. This characterization may be done, for example, through user observation, questionnaires and interviews

or through search engines' logs analysis.

The identification of contextual features that may be useful for HIR is the expected deliverable of this phase.

5.2 Retrieval Framework Definition

After the previous phase's conclusion, it will be necessary to propose an information retrieval framework. This framework incorporates the previoulsy identified contextual features and defines the characteristics of the IR system. Contextual features incorporation may be done at either one of the three levels: the *user interface*, *query operations* and *ranking*. It is possible, for example, to define a method to identify the type of clinical question (e.g. background or foreground question as defined by Hersh [11]) behind a query in order to adjust the relevance score of each result. Or the system may predict the user's expertise level and rank the results accordingly.

Other possible lines of action include: using spelling suggestions and automatic query reformulation; faceted search (allowing the user to navigate hierarchically, choosing the order in which categories are presented); *lookahead* (e.g.: automatic completion of query terms, suggestion of popular terms, results annotations); relevance feedback, summarization (e.g.: aggregating query results in a more consumable form, clustering) and visual presentation of the results (e.g.: tables, charts, tag clouds).

5.3 User Context Capture

The acquisition of contextual data may be done in several ways namely through the analysis of the searcher's or group's characteristics, by the analysis of past interaction processes, by observing properties of contents or even by exploiting temporal properties like frequencies and length of searches. This data may be in the format of log files but it can also be obtained from other systems (e.g.: searches' history, user's bookmarks, email or office applications).

In the end of this phase, the methods to capture user context should be implemented.

5.4 Prototypes Development

In this phase, prototypes based on previous findings and developments will be built. It will be done over standard retrieval models and should integrate previously implemented methods responsible for the collection of contextual features.

Two different strategies may be followed according to the context being explored. The system may explore it (e.g.: through the use of relevance feedback techniques) or can allow searchers to explore it (e.g.: through the adaptation of IR interfaces to present contextual information).

During the implementation phases, existing health thesaurus or ontologies whose relations may be used to improve the retrieval processes, may be used. For example, hierarchical relations can be used do term explosions in a query, synonymous relations may be used by the searcher to express a concept in different words and related relations may be used to remind a searcher of different but related terms [11].

5.5 Evaluation

This phase has the goal to define and setup evaluation experiences able to evaluate the developed prototypes. According to the prototype being evaluated, one or more of the following methods may be selected.

The Cranfield model is still the dominant evaluation model in IR [4] and is based on the use of testbeds (documents, queries and relevances assessments collections). It also includes the use the assessment of retrieval performance through standard IR measures like precision and recall. A possible testbed is the one developed by William Hersh, named OSHUMED[1]. Similarly to what has been done by Liu and Chu [14], other databases may also be adapted to be suitable to the defined experience.

Being an obvious solution to the prototypes' evaluation, the Cranfield method has restrictive assumptions on the cognitive and behavioral features of the IR system's environment [4]. For this reason and for the understandable importance of the environment in contextual IR, it's critical the use of alternative approaches.

The model defined by Borlund [4] to evaluate interactive IR systems is a possible method. In this framework, simulated work tasks and alternative performance measures that take into account the non-binary nature of the relevance assessments are used.

Silva and Favela [19] defined a simulated work task but the evaluation was done through the use of questionnaires (one with questions about the retrieval task and another about experience insights) and through activity monitorization.

Another alternative involves the manipulation of specific context features (e.g. task complexity) as proposed by Bell and Ruthven [2].

The use of real tasks (recording real searches and asking searchers to replay the searches)[7] is another option.

6. ISSUES FOR DISCUSSION

Globally, it would be very positive to my research to have feedback from IR experts on the methodology and ideas here presented. Are there any issues/problems I should be aware of? Are there any suggestions for improvement? Are there other available datasets that may be used in this work?

I expect to have concluded the Health Information Seeking Behavior studies at the date of the Doctoral Consortium. Therefore, it would be an excellent opportunity to discuss their results and to discuss the methods I will have defined to capture contextual features. How should these be implemented and integrated in an IR system?

I'm very interested in discuss evaluation methods and metrics of systems where users play a central role like the ones involved in this work. What are the most common problems in IR experimental setups? How are they usually overcome? Are there more testbeds suitable for the health area?

I'm also interested in seeking suggestions about possible research directions. Are there any research questions relevant to the problem that were not mentioned? Are there any pertinent research studies or literature I should study/follow?

7. ACKNOWLEDGMENTS

This work is partially funded by Fundação para a Ciência e a Tecnologia under the grant SFRH/BD/40982/2007.

[1] http://ir.ohsu.edu/ohsumed/ohsumed.html

8. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, May 1999.

[2] D. J. Bell and I. Ruthven. Searchers's assessments of task complexity for web searching. In *Advances in Information Retrieval, 26th European Conference on IR Research, ECIR 2004*, pages 57–71, 2004.

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

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

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

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

[7] D. Elsweiler and I. Ruthven. Towards task-based personal information management evaluations. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 23–30, New York, NY, USA, 2007. ACM.

[8] A. Göker and H. I. Myrhaug. User context and personalisation. In *ECCBR Workshop on Case Based Reasoning and Personalisation*, 2002.

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

[10] R. Haux, E. Ammenwerth, W. Herzog, and P. Knaup. Health care in the information society - A prognosis for the year 2013. *International journal of medical informatics*, 66(1-3):3–21, November 2002.

[11] W. R. Hersh. *Information Retrieval - A Health and Biomedical Perspective*. Springer, December 2002.

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

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

[14] Z. Liu and W. W. Chu. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, 10(2):173–202, April 2007.

[15] C. T. Lopes. Health information retrieval state of the art. Technical report, Faculdade de Engenharia da Universidade do Porto, July 2008.

[16] P. Lyman and H. R. Varian. How much information. Available from: http://www.sims.berkeley.edu/how-much-info-2003 [cited 2009-02-20], 2003.

[17] D. S. Martins, L. H. Z. Santana, M. Biajiz, A. F. do Prado, and W. L. de Souza. Context-aware information retrieval on a ubiquitous medical learning environment. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 2348–2349, New York, NY, USA, 2008. ACM.

[18] S. M. Maviglia, C. S. Yoon, D. W. Bates, and G. Kuperman. Knowledgelink: Impact of context-sensitive information retrieval on clinicians' information needs. *J Am Med Inform Assoc*, 13(1):67–73, Jan–Feb 2006.

[19] J. M. Silva and J. Favela. Context aware retrieval of health information on the web. In *LA-WEB '06: Proceedings of the Fourth Latin American Web Congress (LA-WEB'06)*, pages 135–146, Washington, DC, USA, 2006. IEEE Computer Society.

[20] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 132–141, New York, NY, USA, 2009. ACM.

[21] X. Yan, D. Song, and X. Li. Concept-based document readability in domain specific information retrieval. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 540–549, New York, NY, USA, 2006. ACM.

APPENDIX

A. STUDENT'S STATEMENT

After having completed the curricular year of the Informatics Engineering Doctoral Program, I'm currently writing my thesis proposal and I expect to finish it near the Doctoral Consortium's date. It is therefore an excellent time to discuss my proposal with field's experts and to receive their comments and suggestions. It's probably the moment where this feedback will be more useful to a successful activities' planning. It will also be an opportunity to meet other doctoral students and to see how they are addressing their problems.

This research has challenges in several sub-areas of Information Retrieval and other related fields like information seeking behavior, user modeling, IR evaluation, user studies, human computer interaction and machine learning. The feedback from experts of some of these areas will help to consolidate the thesis proposal and to better structure the involved techniques. SIGIR is a rare chance to meet and talk with researchers across the several sub-areas of IR.

Finally, being SIGIR one of the best conferences in the field of Information Retrieval (IR), it will be a great place to interact with experienced IR researchers, to meet other students with similar research interests and to attend sessions and presentations. This will certainly be a great help towards the development of a high quality research.

Carla Teixeira Lopes

B. ADVISOR'S STATEMENT

Carla Teixeira Lopes is finishing her 1st year doctoral studies at Faculdade de Engenharia, Universidade do Porto. Her Ph.D. program follows a 5-year Informatics degree and a 2-year Information Management M.Sc.. She has also some experience as a teaching assistant. In the Doctoral Program she is currently (March 2009) completing her Ph.D. proposal which will be evaluated by the steering committee.

I consider that Carla can benefit significantly form attending the SIGIR Doctoral Consortium. The opportunity of presenting and discussing her work in the most representative forum for the IR community is especially relevant for a student coming from a small research group. Moreover, Carla is currently exploring several approaches to her problem; she needs to delimit the topic, to be aware of related ongoing work and to gain a better view of existing research groups. This is the point where she can gain most from the discussion with junior researchers facing similar problems and from the advice of senior researchers on how to further focus her research efforts. The Consortium may also provide opportunities for collaboration with other researchers.

Carla has a very solid background in computer science and a broad view of the problems in information management. She can contribute effectively to the discussion of topics in Information Retrieval, both on her specific area and in those she has explored for selecting her research problem. I am sure that she will be a very active participant in the Doctoral Consortium and in the SIGIR conference in general.

Cristina Ribeiro