Context Features and their use in Information Retrieval

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

Abstract

The context in which a search takes place affects the Information Retrieval (IR) process. It affects the searcher's interaction with the IR system, his expectations and his decisions about the documents he retrieves. Therefore, knowing more about what features are important in a searcher's context and what they are used for, can help design more useful and successful IR systems. This paper has three main contributions. It starts with a literature review on the definition of context and on context taxonomies (1). A systematic representation of context features and uses, based on related work, is then proposed (2) and used in a survey on the use of context features in IR (3). This analysis has concluded that *interaction context* is the most used category of features and *Indexing and Searching* are the tasks where context features are most employed. This work, an initial phase of a PhD research, provides a systematic review of what is being done in the area and proposes a taxonomy for IR.

Keywords: Information Retrieval, Context Features, Context Uses

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 fact, 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 retrieved documents (Ingwersen & Järvelin 2005, Harper & Kelly 2006). A contextualised strategy might allow IR systems to learn and predict what information a searcher needs, learn how and when information should be displayed, present results relating them to previous information and to the tasks the user has been engaged in and decide who else should get the new information.

In the field of Information Retrieval, there is a growing interest in improving the search process towards the user needs and context (Bierig & Göker 2006). An early model that has approached IR from the level of context is the one from Belkin (1980). Later, other authors have also developed models (Ingwersen 1996, Saracevic 1997) in which context is at the center of the IR process. Still in the decade of 1980, another project (Saracevic et al. 1988, Saracevic & Kantor 1988a, Saracevic & Kantor 1988b) was dedicated to the characterization of the elements involved in information seeking and retrieving, such as the cognitive context involved in these processes. More recently, several journals (e.g: Information Processing and Management - 2002, 2008; Information Retrieval - 2007) and conferences have given attention to this topic (e.g: Information Retrieval in Context (IRiX) - 2004, 2005; Information Interaction in Context (IIiX) - 2006, 2008).

While there is consensus that context matters (Cool & Spink 2002, Bierig & Göker 2006), there is no agreement on which context elements influence IR (Cool & Spink 2002). The purpose of this paper is threefold. It starts with a literature review on the definition of context and on proposed context taxonomies (Section 2). A systematic representation of context features and uses, based on related work, is then proposed (Section 3) and used in a survey on the use of context features

in contextual IR (Section 3.2). Section 5 outlines the PhD work that will be built on the current survey and identifies issues for discussion in the Symposium.

2. BACKGROUND AND RELATED WORK

Context is one of the most abused terms in IR, being associated to a large range of ideas (Finkelstein et al. 2002). Brézillon (1999) enumerates twelve different definitions from several authors where the lack of consensus is evident. As Dervin (1997) says, “context has the potential of being virtually anything ... [it is] a kind of container in which the phenomenon resides”. The concept crosses several areas of knowledge from cognitive sciences to engineering. This section reports on definitions in domains related to IR and does not intent to do a thorough review of definitions in other areas. The work of Brézillon (1999) presents a more thorough review of context’s definitions in five areas connected to artificial intelligence.

In the literature, some authors have gone further in the characterization of context, defining contextual taxonomies. These structures facilitate the understanding and exploration of context. Some of the main context taxonomies will also be described in this section.

2.1. Context Definition

According to Dourish (2004), *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 the activity. It consists of implicit attributes that describe the user and the environment in which information activities occur. The second perspective sees context as arising from the activity, from which it can’t be separated.

Dey & Abowd (2000) also do an extensive review on context’s definitions. They propose their own definition that encompasses other authors’ definitions. Context is: “any information that can be used to characterize the situation of entities (e.g. a person, a place or an object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves”. This definition matches the first perspective of Dourish. Other authors give definitions of context that also correspond to the first perspective of Dourish. Göker & Myrhaug (2002) present a short and comprehensive definition: “description of aspects of a situation”, similar to the one from Dey & Abowd (2000). Marchionini (1997) had already defined it as a “setting” that has “physical and conceptual/social components, including whether the task is done in collaboration or alone and the information seekers physical and psychological states”. The first definition proposed by Johnson (2003) also equals context to “situation”.

The second definition of Johnson (2003) goes beyond the enumeration of factors to the specification of the active ingredients in a context, noting that they have predictable effects on processes. In this view, context is defined as a relation between the specific ingredients and the processes, which is closer to the second perspective of Dourish. Similarly, Winograd (2001) says that “something is context because of the way it is used in interpretation”. Sato (2001) defines context as “a pattern of behavior or relations among variables that are outside of the subjects of design manipulation and potentially affect user behavior and system performance”. Ingwersen & Järvelin (2005) say that “actors and their components function act as context to one another in the interaction processes. There are social, organizational, cultural as well as systemic contexts, which evolve over time”.

2.2. Context Taxonomies

Ingwersen & Järvelin (2005) present a nested model of context stratification for IR with six dimensions. **Intra-object structures** refers to context obtained from each document where images are contextual to a surrounding text, paragraphs act as context for their own lines and words. **Inter-object contexts** are concerned with the properties of documents, like references, citations, outlinks and inlinks, that give and take context from other objects. **Interaction/session context** is about the social interaction and interactive IR activities, if the searcher is at the core, or

is about the retrieval session, if the interface is at the core of the taxonomy. **Social, systemic, media, work task, conceptual, emotional contexts** are related to socio-organizational and systemic aspects (like the IT, interface and documents), if the searcher is at the core, or are related to information objects and searching actors, if the interface is at the core. **Economic techno-physical and societal contexts** correspond to the prevailing societal infrastructures. Finally, **historic context** is a temporal form of context that includes all past participating actors' experience.

Dey & Abowd (2000) propose a classification of context information based on the entities in which the context is assessed and on categories of context. They define three entities: **places** like regions of geographical space such as rooms or offices, **people** including individual or groups, co-located or distributed and **things** (e.g. physical objects or software components and artifacts like a computer file). Primary and secondary context characterize these entities. Primary context types are: **identity**, **location**, **status/activity** and **time**. These context types may be used to infer additional pieces of context such as the address of a person by her identity. The latter are designated by secondary context types. In their work they also propose categories for uses of context: presentation of information/services to the user, execution of a service and tagging of context to information for later retrieval.

Göker & Myrhaug (2002) present a context taxonomy in which context elements are divided into five main categories. The **task** category is about what the user is doing, his goals, tasks, activities. The **social** one refers to the social aspects of the user, such as information about friends and family or his role. **Personal** context aggregates mental and physical information about the user such as mood, expertise, disabilities. In the **spatio-temporal** category are included attributes like time and location and the **environmental** context is about user surroundings like things, light, people and information accessed by the user.

Briconsouf & Newman (2007) propose a framework to analyse the use of context in health care applications. Their framework has three main axes to characterize context. The **purpose of use of context** presents the three types of context uses proposed by Dey & Abowd (2000). The second axis, **items for context representation**, identifies three main classes to split items of context into: people, environment and activities . The third axis, **organization of context features** proposes other ways to organize context features such as an hierarchical organization that draws from general to local aspects of context, an organization according to the internal and external dimension of context and an organization according to the usefulness of context (relevant or non relevant for the current action).

Mansourian (2008) has also developed a taxonomy for the contextualization of web search with five main categories. The **web user axis** is divided in feelings, thoughts (attitudes and cognitive style), actions (passive vs active users) and competence. The **search tool** and the **search topic** are two other axis of the taxonomy. The fourth axis, **search situation** is divided in place of search, type of search (work-related or everyday life search), immediacy of search and importance of search. The last axis, **information resources** is split in searchability and accessibility, level of provision (publicly available/restricted access) and level of user-friendliness.

From all the reviewed taxonomies, only the one from Ingwersen & Järvelin (2005) has been made for IR. This is the most exhaustive taxonomy, even though it doesn't propose a classification for uses of context. Only the Dey & Abowd (2000) and Briconsouf & Newman's (2007) taxonomies include this categorization. Göker & Myrhaug's (2002) taxonomy is a well known taxonomy in the field of IR.

3. PROPOSAL OF A CONTEXT TAXONOMY FOR IR

Ingwersen & Järvelin's (2005) taxonomy is the most appropriate to our goals. Yet, it does not covers uses of context. Therefore, it is here proposed a context taxonomy for IR composed of two categorizations, one for the context features potentially useful in a IR system (Figure 1) and other

for possible uses of these features in a IR system (Figure 2). The context features category is a variant of the Ingwersen & Järvelin's (2005) taxonomy.

In this proposal, context is considered an interactional problem, as defined by Dourish (2004). It is considered that it does not only deal with the environmental features surrounding the user and its activities, but also concerns the interaction in other tasks and situations in similar domains. Context evolves over time and users' context can change each time a new search is made, a new set of results is reviewed or a new document is viewed (Harper & Kelly 2006). Therefore, "it arises from and is sustained by the activity itself" (Dourish 2004).

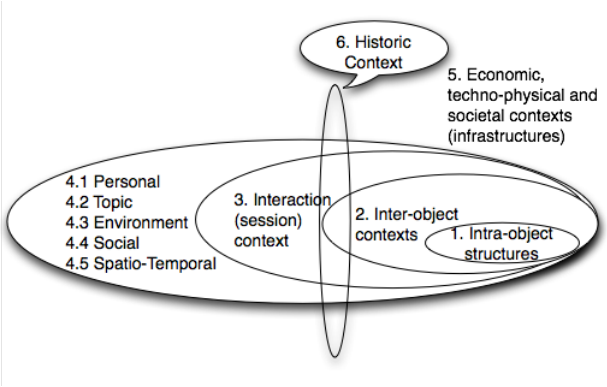

FIGURE 1: Taxonomy for Context Features - variant of (Ingwersen & Järvelin 2005)

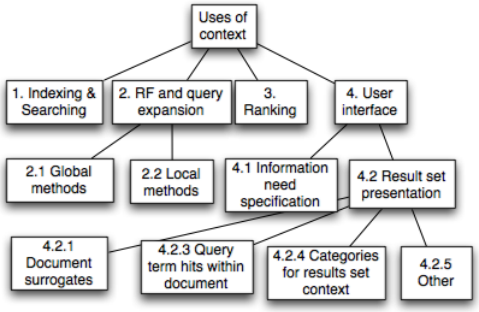

FIGURE 2: Taxonomy for Uses of Context

3.1. Context Features

The proposed taxonomy is similar to the one proposed by (Ingwersen & Järvelin 2005). It dffers in its fourth dimension that is mainly modeled by the context's categories defined by (Göker & Myrhaug 2002). The adoption of the model of Ingwersen & Järvelin is justified by the presence of several relevant dimensions in the IR domain. This model's fourth dimension includes the majority of the other taxonomies' categories. The option for integrating the Göker & Myrhaug's (2002) taxonomy is explained by its comprehensiveness and by its clear and logical partition of context features in categories.

The model is defined by 6 dimensions. Dimensions 1 and 2 are related to the intra/inter-object contexts. The first is related to the documents' structure and content that may act as context. The second is about documents' properties that relate them with other documents. The information resources' category of the Mansourian's (2008) taxonomy fall in the intra-object context category.

In Ingwersen & Järvelin's (2005) model, the third dimension may be approached in two different ways according to who is at the core of the model: the user or the interface. As the proposed taxonomy is centred on the user, this category is about all the social interaction and activities that occur inside the IR session: "what the persons (actors) are doing [...] can be described with explicit goals, tasks, actions, activities, or events. [...] can include other persons tasks (that are within the situation)" as defined by Göker & Myrhaug. Task context can also be characterized by variables like endurance, frequency and stage (Kelly 2006). This dimension contains the activities category in the Briconsouf & Newman's (2007) taxonomy and a part of the status/activity category in the Dey & Abowd's (2000) taxonomy.

The fourth dimension joins the other four categories of the Göker & Myrhaug's (2002) taxonomy. The **personal context** contains the physiological (e.g. "pulse, blood pressure, weight, glucose level, retinal pattern, and hair colour") and the mental context (e.g. "mood, expertise, angriness, and stress"). The **topic context** has information about the persistence and familiarity of the user with the topic (Kelly 2006) and may also contain information about its nature (work or non-work related; fact or subject search) (Mansourian 2008). **Environment context** captures the entities that surrounds the user such as things, services, temperature, light, humidity, noise, persons,

physical constraints (e.g. amount of time, physical accessibility, comfort, cost) and surrounding information. **Social context** has information about "friends, neutrals, enemies, neighbours, co-workers, and relatives for instance". It also includes the roles played by the user, his status in these roles, the tasks he can perform in each role and the various sub-roles he can have. **Spatio-temporal context** describes aspects such as time, location, direction, speed, shape, track, place, clothes of the user, this is, the spatial extension of the environment and the things in it.

The two last dimensions are **economic techno-physical and societal contexts** and **historic context**. The first is more global than the environment context in the fourth dimension. It can include actual and global aspects like the H1N1 flu or the economic crisis. The sixth dimension involves all user's past actions.

3.2. Context uses

In this section are presented categories of uses of context in IR. From the authors of the reviewed context taxonomies, only Dey & Abowd (2000) has propose such an organization (which was later included in the Briconsouf & Newman's (2007) taxonomy). Their organization has three categories: presentation of information and services to a user, automatic execution of a service and tagging of context to information for later retrieval.

With the Dey & Abowd's (2000) categories in mind and with IR as this work's focus, the proposed top-level categories of uses of context in IR are: **indexing & searching**, **relevance feedback (RF) and query expansion**, **ranking**, **user interface**. These categories are components of an IR system were context may be used. The proximity of techniques used in the index construction and searching phases, stimulated their fusion in a single category. The proposed categories also map perfectly well to the categories defined by Dey & Abowd: the indexing & searching fits in the tagging category; the RF and query expansion may fit in the presentation of information and services (e.g. relevance feedback) or automatic execution of a service (e.g. implicit relevance feedback); the ranking fits in automatic execution of a service; and the user interface fits in the presentation of information and services.

The RF and query expansion category involves the processes of query refinement by the system, either fully automatically or with the help of the user. As defined by Manning et al. (2008), this category is divided in global and local methods. **Global methods** include query expansion/reformulation based on collection-independent knowledge structures (Efthimiadis 1996) like domain-specific thesaurus or general-purpose thesaurus (e.g.: WordNet), query expansion via automatic thesaurus generation and techniques like spelling correction. **Local methods**, like relevance feedback, pseudo relevance feedback and implicit relevance feedback, adjust the query with information from the documents that belong to the result set of the initial query. In relevance feedback the user marks returned documents as relevant or non-relevant and the system builds a better representation of the information need based on his feedback (Manning et al. 2008). Pseudo relevance feedback assumes the k ranked documents as relevant and implicit relevance feedback uses indirect sources of relevance.

The user interface category is also divided in two subcategories: the interface associated with the specification of the user's information need and the presentation of the result set. This last category is also divided in document surrogates (e.g. snippet - short summary of the document), query term hits within document (e.g. keyword-in-context snippets), categories for results set context and other type of strategies.

4. USE OF CONTEXT FEATURES IN IR

The proposed taxonomy was the basis for the analysis of a sample of contextual IR research papers. This sample is composed of 25 papers whose references are available at `http://www.carlalopes.com/papers_sample.pdf`. Papers' selection was made from a set of papers classified with the tag context (`http://www.citeulike.org/tag/context`) in *CiteULike*, a social web

service for management of bibliographic references. In this list, papers related to IR, published in 2008, that made use of context features were included in our sample.

Each paper was examined towards the identification of: context definition adopted, context taxonomy exploited, context features used in the experience and their specific use. Only four papers introduced the adopted context definition and only one presented the underlying context taxonomy. Figure 3 has two pie charts where the left one shows the proportion of papers using each context feature's category and the right one shows the proportion of implemented context uses. In these graphs, each context feature (CF) and context use (CU) is represented by the numbers given in Figures 1 and 2.

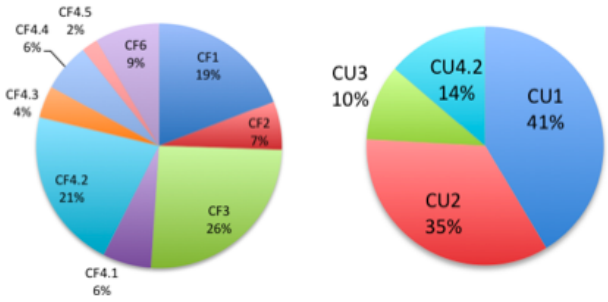

FIGURE 3: Context features and its uses in the set of analysed papers

The most used context features are the *interaction (session) context*, the *topic* and the *intra-object structures*. Interaction features range from desktop and web user behavior to users' tasks, actions and submitted queries. The context features in the topic category are as diverse as TREC topics' descriptions, context documents, domain thesaurus/ontologies and conceptual maps. *Indexing and searching* is the IR system's component where more papers employ context features, followed by the *RF and query expansion* with 35% of the papers.

5. FUTURE WORK AND ISSUES FOR DISCUSSION

This work is an initial phase of a PhD research that seeks to study how context features surrounding health information seeking and retrieval can affect the use of Health IR (HIR) systems and to apply these features in the improvement of these systems. The next step involves conducting an Health Information Seeking Behavior study to find the context attributes that matter for HIR applications. It will then be necessary to find ways to capture the identified context features and to define strategies to improve HIR involving the identified context elements.

Several issues are relevant for discussion that will, undoubtedly, be of great value to this PhD research. It would be interesting to discuss: ways to exploit shared contexts and contexts over time; evaluation methods and metrics of systems where users play a central role; common problems in IR experimental setups; ways to overcome these problems; testbeds suitable for the health area; envisioned research directions and pertinent research studies or literature to study.

6. ACKNOWLEDGEMENTS

I would like to acknowledge Cristina Ribeiro for reading and giving suggestions to this paper. This work is partially funded by Fundação para a Ciência e a Tecnologia under the grant SFRH/BD/40982/2007.

References

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval, *Canadian Journal of Information Science* (5): 133–143.

Bierig, R. & Göker, A. (2006). Time, location and interest: an empirical and user-centred study, *IliX:*

Proceedings of the 1st international conference on Information interaction in context, ACM, New York, NY, USA, pp. 79–87.

Brézillon, P. (1999). Context in problem solving: a survey, *Knowl. Eng. Rev.* **14**(1): 47–80.

Briconsouf, N. & Newman, C. (2007). Context awareness in health care: A review, *International Journal of Medical Informatics* **76**(1): 2–12.

Cool, C. & Spink, A. (2002). Issues of context in information retrieval (ir): an introduction to the special issue, *Information Processing & Management* **38**(5): 605–611.

Dervin, B. (1997). Given a context by any other name: methodological tools for taming the unruly beast, *ISIC '96: Proceedings of an international conference on Information seeking in context*, Taylor Graham Publishing, London, UK, UK, pp. 13–38.

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

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

Efthimiadis, E. N. (1996). Query expansion, *Annual Review of Information Systems and Technology (ARIST)* **31**: 121–187.

Finkelstein, L. E. V., Gabrilovich, E., Matias, Y., Rivlin, E. H. U. D., Solan, Z. A. C. H., Wolfman, G. A. D. I. & Ruppin, E. (2002). Placing search in context: the concept revisited, *ACM Trans. Inf. Syst.* **20**(1): 116–131.

Göker, A. & Myrhaug, H. I. (2002). User context and personalisation, *ECCBR Workshop on Case Based Reasoning and Personalisation*.

Harper, D. J. & Kelly, D. (2006). Contextual relevance feedback, *IliX: Proceedings of the 1st international conference on Information interaction in context*, ACM Press, New York, NY, USA, pp. 129–137.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory, *Journal of Documentation* **52**(1): 3–50.

Ingwersen, P. & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*, 1 edn, Springer.

Johnson, J. (2003). On contexts of information seeking, *Information Processing & Management* **39**(5): 735–760.

Kelly, D. (2006). Measuring online information seeking context, part 1: Background and method, *Journal of the American Society for Information Science and Technology* **57**(13): 1729–1739.

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

Mansourian, Y. (2008). Contextualization of web searching: a grounded theory approach, *The Electronic Library* **26**(2): 202–214.

Marchionini, G. (1997). *Information Seeking in Electronic Environments (Cambridge Series on Human-Computer Interaction)*, Cambridge University Press.

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications, *Proceedings of the American Society for Information Science*, Vol. 34, pp. 313–327.

Saracevic, T. & Kantor (1988a). A study of information seeking and retrieving. ii. users, questions and effectiveness. *Journal of the American Society for Information Science* **3**(39): 177–196.

Saracevic, T. & Kantor (1988b). A study of information seeking and retrieving. iii. searchers, searches and overlap, *Journal of the American Society for Information Science* **3**(39): 197–216.

Saracevic, T., Kantor, P., Chamis, A. Y. & Trivison, D. (1988). A study of information seeking and retrieving. i. background and methodology, *Journal of the American Society for Information Science* **3**(39): 161–176.

Sato, K. (2001). Context sensitive interactive systems design: a framework for representations of contexts, *Hum. Comput. Interact.*, Vol. 2, pp. 229–241.

Winograd, T. (2001). Architectures for context, *Hum.-Comput. Interact.* **16**(2): 401–419.