Proceedings of the World Conference "E-learn 2004", Washington, DC, USA, Nov. 1-5, 2004

How the Notion of Context can be Useful to Search Tools

Bich-Liên Doan
Supélec, Plateau de Moulon, 3 rue Joliot Curie, 92192 Gif/Yvette, France.
E-mail : Bich-Lien.Doan@supelec.fr

Patrick Brézillon
LIP6, University Paris 6, 8 rue du Capitaine Scott, 75015 Paris, France
E-mail : Patrick.Brezillon@lip6.fr

Abstract

In this article, we suggest some recommendations for helping the search tools to decrease noise and silence. Taking account different contextual spaces (user, search tool, document and interactions between the user and the search tool), we explain how a search tool may provide customized responses within a "specific user context".
***Keyword**: Context, Information Retrieval on the Web*

1. Introduction

Experiments by [1] show that most of user's requests contain 2 or 3 terms. So few numbers of terms leads necessarily to noise and silence in the responses given by search tools. This is due to several factors that include, among others, the strong complexity and heterogeneity of the information available on the Web, the implicit intentions of the user, the environment the user is attached to and the system itself. Such factors constrain the search without intervening in it explicitly. This is what is called context [2]. One idea to improve the efficiency of the information retrieval systems (IRSs) is to make explicit the context the query belongs to. The context is linked not only to the terms composing the query, it is linked to the user-profile, to the system itself and finally it has to be determined dynamically during the interactions between the system and the user. Informally, we define the context in Information Retrieval on the Web as the sum of the following contexts:

- The context of the user and its environment,
- The context of the information provided to the IRS (documents and authors),
- The context of the IRS,
- The context of the interactions between the user and the IRS,

Context may be explicit or implicit. Explicit context is the information we can retrieve directly under the form of physical information. This is also some contextual knowledge that is introduced explicitly in the search process. Implicit context is rather inferred from either the explicit context or from the interactions between the system and the user. In this article, we are solely interested in the explicit context.

Some widely-used means to evaluate the efficiency of IRSs is to compute the relevance of the queries. Generally, some test collections containing documents and queries are provided with a set of relevant responses manually given by experts (ex: TREC [4]). The relevance of the responses given by each IRS is then evaluated according to that predefined set of responses. Because of the diversity of information sources and users on the Web, and because of the continuously changing of the information available on the Web, it has become difficult to build a really significant reference and stable test collection.

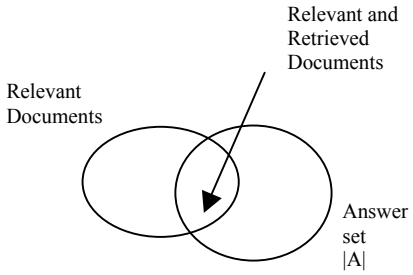

Figure 1: Evaluation model of an IRS

Figure 1 shows a classical model for evaluating an IRS, thanks to the precision and the recall measure. The first oval represents the set of relevant responses whereas the second oval represents the set of responses retrieved by the IRS. The oval intersection is the set of relevant responses retrieved by the IRS.

Our assumption is that each user has her/his own contextual space (an individual context), i.e. the set of relevant responses differs from one user to another according to his profile. Contrary to the Figure 1, in Figure 2 there are different sets of relevant responses taking account different contexts, such as the user's profile, the description of documents and the context of the search tool:

On the following sections, we will take the same example of a query: "context + information + retrieval" in order to retrieve documents more

particularly in the "context" or in the "information retrieval" domain.

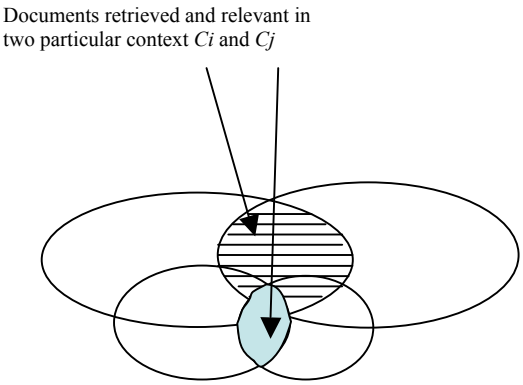

Figure 2: Context dependency of relevance

2. Explicit Context

2.1 User profile context :

A user's profile is often encoded using a weighted vector according to the TFIDF. This profile represents the user's preferences and it is progressively filled by the mechanism of the relevance feedback, where the asker reacts to the system by saying whether or not the document is relevant. On the next iteration, the filtering task simply indicates to the user which documents may be relevant to him according his profile. In Windows NT, a user profile is a record of user-specific data that define the user's working environment. The record can include display settings, application settings, and network connections. What the user sees on his computer screen, as well as what files, applications and directories they have access to, is determined by how the network administrator has set up the user's profile.
A user's context contains a user profile. In the following, we catalogue the information that is part of the user profile. A user may be specialized or not in the area of the information he wants to obtain. For example a user is specialized in the "context" area or in the "information retrieval" area. The personal bibliography, the personal Web site, the curriculum vitae, the log files, the bookmarks and the documents of his directories can fill the user profile. The way a user organises his information (tree of directories or flat organization) reveals the structure he is used to tidying his files. The community of researchers, the organism the user works in, the persons known, the network of accountancies are elements of the profile which will help to determine which documents the user wants to retrieve.
The social context of the user is the second part of the user's context and depends on the location, if he works at home or at work. Some material constraints such as the speed of transmission, the volume of the information to download, the URL location, the origin of the pages (from a commercial or an academic institute site), the exploitation of the cache, the availability of the user may influence the selection of an URL.
We give an example of a user's context described with the XML syntax:

```
<user_context>
<user_profile>
<identity>Bich-Liên Doan</identity>
<research_area>information
retrieval</research_area>
<organism>Supelec</organism>
<file_organisation>tree</file_organisation>
<personal_biblio>http://www.supelec.fr/si/pers
onal/bld.html</personal_biblio>
<external_bibio>
http://www.supelec.fr/si/external/bld.html</ext
ernal_biblio>
<accountancies><list><item1>Patrick
Brézillon><item1>
<item2>JT<item2></list></accountancies>
</user_profile>
<social_context>
<location>Supelec</location>
<transmission_speed>high</transmission_spee
d>
<domaine_site>.org, .edu</domaine_site>
<ttl>10 minutes</ttl>
<cache>yes</cache>
<volume>"">1Mo", "site>500pages"</volume>
</social_context>
</user_context>
```

In this example, the "domain_site" and the "transmission_speed" help the search tool to control the origin of the documents and to enable voluminous documents to be downloaded (by calculating how many times it will take to download documents).
To precise the semantics of the field, the use of an XML-schema is mandatory. We can use inclusion or inheritance relations between contexts. For example, a user context may include a context defined inside a community. This community may be the institution he works for (for example Supelec) or a research community (for example Information Retrieval). A user context may inherit from another user context.

2.2 IRS context :

An IRS provides a context which depends on the collected resources, the indexing algorithm, the matching algorithm, the query language and the display of results. We can distinguish three types of search tools on the Web: the classical search tools (Google), the metasearchers (metacrawler), and the directories (Yahoo!). A search tool first makes a corpus, formed by the resources the

robot may collect from the Internet which is size-limited by how many pages on the Web it covers (for example Google collects 20% of the resources on the Web). The main feature of the Web is its dynamicity. The volume of information available on the Web increases exponentially, and every month, 40% of the Web changes. But 30% of the information on the Web is duplicated.
A second feature is the indexing of the corpus. The indexing language is the indexing context. Indexing is the process of associating one or more keywords with each document.
The result of a query is featured by the number of references given as responses and by the output of the results: summary of the page, two first lines with keywords underlined, the title of the page, the URL of the page and the site this page belongs to. The similar pages, a thematical hierarchy or the organisation of the responses, the format of the files. The user has different uses of the responses, he may want to store the file, to copy and paste it or to read it only. He can choose a search tool and get used of the query language and the display mode.
The filtering of the query or of the output provides some restriction upon the language, the date, the format of page…
The user can select one part or several parts of the page. The language of the site can be automatically chosen by the search tool (google.fr or google.com)

```
<IRS_context>
<corpus>
<volume> 20% of the Web</volume>
<type>text, image </type>
<collect_frequency>1%  every
day</collect_frequency>
</corpus>
<indexing>
<type> text</type>
<fields>free text, metadata, tags, type</fields>
<algorithm>thesaurus, TFIDF, weighted
tags</algorithm>
</indexing>
</IRS_context>
```

2.3 Document context :

The notion of document context has been used in [3]. A document is defined as a physical resource found in response to a query, identified by an URL. A hypertext system is composed of several documents linked with hypertext links. One page belongs to a site, the others pages of the site provide some context to that page. The organisation of the site thought by the author may not interest the user, who has another vision of this organisation. He may only be interested by some part of the information even if it can be easier for him to detect the context thanks to this organization.
This is an example of a document described by its author:

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<educational_resources>
<General>
  <title><langstring xml:lang=”en”>”Computer
science history, image of the
  Cray CDC 6600”</langstring></title>

<identifier><entry>”http://www.bld.fr/eduReso
urces/”</entry></identifier>
  <keyword>”course, computer science, history,
Cray”</keyword>
  <structure>”atomic”</structure>
  <agregationlevel>”1”</agregationlevel>
</General>
<LifeCycle>
  <contribute><entity>Bich-Liên
Doan</entity><date>”2000
september”</date>
  </contribute>
</LifeCycle>
<Educational>

<intendedenduserrole>”learner”</intendedendu
serrole>
  <context>”higher education”</context>
</Educational>
  <Technical><format>”img/bmp”</format>
</Technical>
</educational_resources>
```

Typically an author produces text that is a balance between what he wants to say and what he thinks the audience wants to hear. A textual corpus contains documents written by different authors for many different purposes. A document is about a topic, we have to understand both. Indexing is the process of associating one or more keywords with each document. What is the context in which these keywords are going to be interpreted? Who's the audience? The real issue is not only to describe a document but to distinguish it from others.

2.4 User/IRS interaction :

Queries are generated as an attempt by users to express their information need. An initial query begins the dialog; the search engine's response provides clues to the user about directions to pursue next; these are expressed as another query. There are alternatives of user-search engine exchanges. If users click on documents they like, the search engine can, by itself, form a new query that focuses on those keywords that are especially associated with these documents. There are many things we can learn from the entire query session.
All documents have equal about-ness. The smallest unit of text with appreciable about-ness is the paragraph. All manner of longer documents are constructed out of basic paragraph atoms. A word, even a sentence does not by itself provide enough context for any question to be answered or found out about.

Queries will often refer to aspects of both free text and metadata.

3. Conclusion

In this paper, we discussed about several expressions of contexts in order to enable search tool to focus on the right user need. Using a vector model, we will test the responses in a context with a TREC base. We are building an additional index enriched with the contextual information, which will be used in the query processing. Our work allows to point out the different types of context that intervene during a session, the relationships between them and thus the movement of information (and eventually its transformation by a misinterpretation) across contexts. The next step in our research in context-based information retrieval is to design and develop an efficient model of context tailored to information retrieval.

4. Bibiliography

[1] B.J. Jansen, A. Spink, J. Bateman, T. Saracevic, "Real Life Information Retrieval: A study of user queries on the Web". *SIGIR-Forum*, Spring 1998. vol32 No 1. pp. 5-17.

[2] P. Brézillon, "Context in problem solving: A survey". *The Knowledge Engineering Review*, *vol. 14*(1), 1999, 1-34.

[3] B-L. Doan, Y. Bourda, H. Delebecque, "Using Links between Educational Resources in the Indexing and Query Step", *World Conference on Educational Multimedia & Telecommunications 2003 (Ed-MEDIA 2003)*, Honolulu, Hawaii, USA, June 23-28, 2003, pp 1235-1238.

[4] http://trec.nist.gov/