# Contextual Retrieval of Documents in Integrated Data Providers

Renan Rodrigues de Oliveira [1] and Cedric Luiz de Carvalho [1]

[1] Institute of Informatics – Federal University of Goiás
PO Box 131 – 74.001-970 – Goiânia – Brazil

{renan, cedric}@inf.ufg.br

**Abstract.** The system described in this work addresses the integration of a set of digital libraries, repositories and other data providers, integrated through the OAI-PMH protocol. Subsequently, this integrated repository is used for contextualized documents recovery, considering the Portuguese language. The definition of contexts has been implemented using ontologies and content analysis of articles in wiki environments. The result presented by developed system is a ranking of documents ordered by a higher degree of importance with respect to the query provided by a user and a particular domain of knowledge, which is specified by an ontology.

**Keywords:** OAI-PMH, Data Integration, Ontologies, Information Retrieval

## 1 Introduction

The system was developed in two modules. The first is responsible for the integration of data providers and the second for retrieving documents. Access can be made by calling web services or through a web interface.

Queries are submitted in the form of a string of keywords (terms). This sequence of terms goes through a stage of preprocessing. Then, the query is expanded with the addition of new terms. The new list of terms is placed in the form of a boolean expression and submitted to a database formed by the integration of various data providers. This integration is done through the OAI-PMH protocol [3].

Environments wiki [4] are used as a knowledge base to assist the extraction of relevant terms. The relevance of these terms is determined from the domain knowledge, specified by an ontology [2]. To calculate the similarity between the query and documents retrieved, we used the Jaccard Coefficient [1].

## 2 Results

Table 1 shows an example of a query with the words “Colisão deTrânsito” (Traffic Colision), within the domain of knowledge “Trânsito” (Traffic). Figure 1 shows the distribution of the number of documents retrieved in response to the query in Table 1. This table shows the range of values of similarity of documents according to the

query and knowledge domain specified by the ontology. In total, 74 documents were retrieved.

**Table 1.** Query on "Colisão de Trânsito" (Traffic Colision), considering "Trânsito" (Traffic) as the knowledge domain.

| ***Query*** | ***Domain of Knowledge*** |
|---|---|
| Colisão de Trânsito | Trânsito |
| ***Query expansion*** | |
| colisão, acidente, trânsito, tráfego | |
| ***Articles Analyzed of Environments Wiki*** | |
| http://pt.wikipedia.org/wiki/Acidente_de_automóvel<br>http://pt.wikipedia.org/wiki/Trânsito | |
| ***Selection terms of Environments Wiki*** | |
| colisão, acidente, motoristas, condutor, trânsito, tráfego, transportes, atropelamento, automóvel, carro, feridos, pedestres, estrada, código, ctb, ruas | |

In Figure 1, for example, to a threshold value greater than or equal to 0.4, 14 of 74 documents can be considered more relevant. This quantity is obtained by adding up the values found in the third, fourth and fifth columns of this figure.

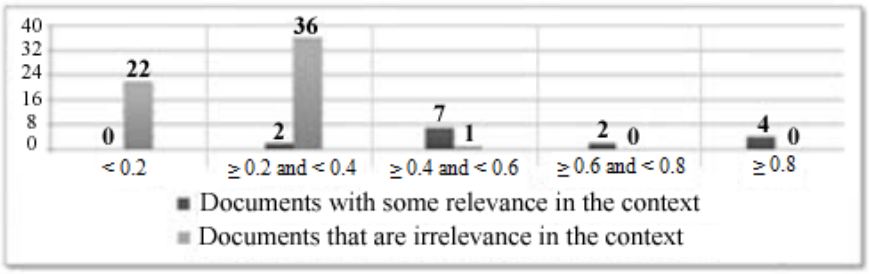


**Fig. 1.** Documents retrieved in response to the query "Colisão de Trânsito" (Traffic Colision), considering "Trânsito" (Traffic) as the knowledge domain.

As shown in Figure 1, the system can separate the documents deemed relevant with respect to the query and a particular field of knowledge. However, note that some documents had some relevance to a low value of similarity. This behavior is primarily due to the low quality of metadata for these documents because they are poorly filled or with very concise descriptions.

## References


1. C. D. Manning, An Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 2009.
2. D. Allemang, J. Hendler, Semantic Web for the Working Ontologist – Effective Modeling in RDFS and OWL, Morgan Kaufmann, 2008.
3. OAI-PMH, http://www.openarchives.org/pmh.
4. P. Shönhofen, et al., Cross-Language Retrieval with Wikipedia, Proceedings of the Workshop of the Cross-Language Evaluation Forum (CLEF), 2007.