# Information Retrieval in Collaborative Environments

**PhD Proposal for the MAP-i Program, 2013-14 Edition**

## 1. Advisor

**Sérgio Nunes** (sergio.nunes@fe.up.pt), Senior Researcher at INESC TEC and Assistant Professor at the Department of Informatics Engineering, Faculty of Engineering (FEUP), University of Porto. http://web.fe.up.pt/~ssn

## 2. Description

One of the central motivations for the initial development of the World Wide Web was to foster the sharing of information between researchers. Since its first version, launched more than 20 years ago, the World Wide Web has largely surpassed all expectations to become a global communication medium with unique characteristics.

A particularly relevant aspect of modern information networks is the promotion of collaboration between users to produce shared information resources, both public or private. The Wikipedia is a prime example of a public resource built and maintained by the collaboration of millions of users. Within organizations, collaborative tools, such as wikis or other document sharing tools, are commonly used to organize and develop work.

Information Retrieval (IR) is the Informatics field primarily focused on all problems and challenges related to information storage and access. It is our belief that information retrieval over dynamic documents, such as those produced collaboratively, can be substantially improved by looking at the dynamic features of those documents. Instead of only considering the last version of a dynamic document, IR systems should incorporate time-dependent measures in their algorithms.

**With this proposal we intent to address this opportunity and investigate how the evolution of documents and collections can be used to improve information access to the current version of these documents.**

## 3. Research Goals

This proposal is centered on studying and developing new techniques to improve information access within the context of collaborative documents. We believe that the temporal information embedded in these documents will be a decisive factor in this process.

Our intuition is that by using the temporal information attached to each collaborative document we can significantly improve retrieval performance in typical IR tasks — e.g. keyword extraction, document summarization, document raking. Furthermore, we believe that by exposing the temporal dimension of documents to the users, new information needs will arise. We have already shown that the use of temporal information introduces significant improvements to retrieval operations (see Previous Work). The main goal is to build upon these results and further improve retrieval tasks.

## 4. Previous Work

Sérgio Nunes, Cristina Ribeiro, and Gabriel David. *Term weighting based on document revision history.* Journal of the American Society of Information Science and Technology (JASIST), 2011. Pre-print available at http://web.fe.up.pt/~ssn/pubs/nunes11-jasist.pdf

Sérgio Nunes, Cristina Ribeiro, Gabriel David. *Term Frequency Dynamics in Collaborative Articles.* 10th ACM Symposium on Document Engineering (DocEng2010). Setembro 2010. Manchester, Reino Unido. http://doi.acm.org/10.1145/1860559.1860620

Sérgio Nunes, Cristina Ribeiro,Gabriel David. *WikiChanges - Exposing Wikipedia Revision Activity.* 4th International Symposium on Wikis (WikiSym 2008), September 2008. Porto, Portugal. http://dx.doi.org/10.1145/1822258.1822292