Inf Retrieval (2006) 9: 71–93
DOI 10.1007/s10791-005-5722-4

A merging strategy proposal: The 2-step retrieval status value method

Fernando Martínez-Santiago · L. Alfonso Ureña-López · Maite Martín-Valdivia

Received: December 9, 2003 / Revised: December 10, 2004 / Accepted: January 12, 2005

Abstract A usual strategy to implement CLIR (Cross-Language Information Retrieval) systems is the so-called query translation approach. The user query is translated for each language present in the multilingual collection in order to compute an independent monolingual information retrieval process per language. Thus, this approach divides documents according to language. In this way, we obtain as many different collections as languages. After searching in these corpora and obtaining a result list per language, we must merge them in order to provide a single list of retrieved articles.

In this paper, we propose an approach to obtain a single list of relevant documents for CLIR systems driven by query translation. This approach, which we call 2-step RSV (RSV: Retrieval Status Value), is based on the re-indexing of the retrieval documents according to the query vocabulary, and it performs noticeably better than traditional methods.

The proposed method requires query vocabulary alignment: given a word for a given query, we must know the translation or translations to the other languages. Because this is not always possible, we have researched on a mixed model. This mixed model is applied in order to deal with queries with partial word-level alignment. The results prove that even in this scenario, 2-step RSV performs better than traditional merging methods.

Keywords CLIR · Merging strategies · Pseudo-relevance feedback · 2-step RSV · Mixed 2-step RSV

1. Introduction

The typical CLIR requirement is for the user to input a free form query, usually a brief description of a topic, into a search or retrieval engine which returns a list, in ranked order, of documents or web pages that are relevant to the topic. The search engine matches the terms in the query to indexed terms, usually keywords previously derived from the

F. Martínez-Santiago · L.A. Ureña-López · M. Martín-Valdivia
Department of Computer Science, University of Jaén, Jaén, Spain
e-mail: {dofer, laurena, maite}@ujaen.es