

ÁP DỤNG HỌC MÁY DỰA TRÊN LẬP TRÌNH DI TRUYỀN TRONG TÌM KIẾM WEB XUYÊN NGỮ

LEARNING TO RANK BASED ON GENETIC PROGRAMMING FOR CROSS-LANGUAGE WEB SEARCH

Tác giả: [Lâm Tùng Giang*](#), [Võ Trung Hùng](#), [Huỳnh Công Pháp](#)

Tóm tắt bằng tiếng Việt:

Hầu hết các nghiên cứu trong lĩnh vực truy vấn thông tin xuyên ngữ giới hạn xem xét các tài liệu văn bản và chú trọng xử lý vấn đề dịch thuật. Trong bài báo này, chúng tôi đề xuất áp dụng học xếp hạng dựa trên kỹ thuật lập trình di truyền nhằm tăng hiệu quả của hệ thống tìm kiếm web xuyên ngữ. Cụ thể, chúng tôi đề xuất 2 phương pháp xây dựng các hàm xếp hạng mới dưới dạng tổ hợp tuyến tính của các hàm xếp hạng cơ sở. Đồng thời, chúng tôi cũng đề xuất 2 mô hình xếp hạng lân cận, ứng dụng trong truy vấn xuyên ngữ. Trong thí nghiệm với một hệ thống tìm kiếm web xuyên ngữ Việt-Anh, điểm số MAP trung bình sử dụng phương pháp kiểm định 5-thuật của các mô hình đề xuất là 0,4640 và 0,4585, vượt trội so với điểm MAP 0,3742 của cấu hình cơ sở - sử dụng bản dịch thủ công.

Từ khóa: tìm kiếm xuyên ngữ; lân cận; xếp hạng lại; học xếp hạng; lập trình di truyền; tìm kiếm web

Tóm tắt bằng tiếng Anh:

Most studies in the field of Cross-Language Information Retrieval consider the documents as plain texts and mainly focus on translation problems. In this article, we follow the learning to rank approach based on Genetic Programming to improve ranking performance of a cross-language web search system. We also introduce 2 proximity models, applied in cross-language information retrieval. We propose linear combinations of weak rankers for re-ranking the retrieved documents. In our experiment with a Vietnamese-English cross-language web search system, the performance measured by the MAP score and reported by a 5-fold cross validation of proposed models is 0.4640 and 0.4585. These results outperform the MAP score of 0.3742 given by the baseline configuration, using the manual translation.

Key words: Cross Language Information Retrieval (CLIR); proximity; re-ranking; learning to rank; Genetic Programming; web search