Filtering for Improving the Geographic Information Search

José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, and L.A. Ureña-López

SINAI Research Group, Computer Science Department, University of Jaén, Spain
{jmperea,magc,mgarcia,laurena}@ujaen.es

Abstract. This paper describes the GEOUJA System, a Geographical Information Retrieval (GIR) system submitted by the SINAI group of the University of Jaén in GeoCLEF 2007. The objective of our system is to filter the documents retrieved from an *information retrieval* (IR) subsystem, given a multilingual statement describing a spatial user need. The results of the experiments show that the new heuristics and rules applied in the *geo-relation validator* module improve the general precision of our system. The increasing of the number of documents retrieved by the *information retrieval* subsystem also improves the final results.

1 Introduction

This paper describes the second participation of the SINAI[1] research group of the University of Jaén in GeoCLEF 2007[1]. In GeoCLEF 2006 we studied the behavior of query expansion[2]. The results showed us that the expansion of topics did not improve the baseline case. However, the results obtained in GeoCLEF 2007 make clear that filtering documents retrieved increases the precision and the recall of a geographical information search.

In GeoCLEF 2007 the results obtained showed that the heuristics applied were quite restrictive. In the post experiments presented in this paper we have defined additional rules and heuristics, less restrictive than used in the official task. In the *geo-relation validator* module, the most important subsystem in our architecture, we have eliminated the heuristic that considered entities appearing in query without an associated *geo-relation*. In addition, the number of retrieved documents by the *information retrieval* subsystem has been increased too, in order to provide a larger variety of documents to be checked by the *geo-relation validator* subsystem.

The next section describes the system overview. Then, each module of the system is explained in the section 3. In the section 4, experiments and results are described. Finally, the conclusions and future work are expounded.

[1] http://sinai.ujaen.es

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 823–829, 2008.

documents validated by the GR validator subsystem and the queries will be run against the new index.

Acknowledgments

This work has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id_514 granted by University of Jaén.

References

1. Perea-Ortega, J.M., García-Cumbreras, M.A., García-Vega, M., Montejo-Ráez, A.: GEOUJA System. University of Jaén at GEOCLEF 2007. In: Working Notes of the Cross Language Evaluation Forum (CLEF 2007), p. 52 (2007)
2. García-Vega, M., García-Cumbreras, M.A., Ureña-López, L., Perea-Ortega, J.M.: GEOUJA System. The first participation of the University of Jaén at GEOCLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
3. Porter, M.F.: An algorithm for suffix stripping. Program 14, 130–137 (1980)
4. García-Cumbreras, M.A., Ureña-López, L.A., Santiago, F.M., Perea-Ortega, J.M.: BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. LNCS. Springer, Heidelberg (2007)
5. Robertson, S., Walker, S.: Okapi-Keenbow at TREC-8. In: Proceedings of the 8th Text Retrieval Conference TREC-8, pp. 151–162. NIST Special Publication 500-246 (1999)
6. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic query expansion using smart: Trec 3. In: Proceedings of TREC3, pp. 69–80. NIST, Gaithesburg (1995)