

GeoTextMESS: Result Fusion with Fuzzy Borda Ranking in Geographical Information Retrieval

Davide Buscaldi¹, José M. Perea Ortega², Paolo Rosso¹, L. Alfonso Ureña López², Daniel Ferrés³, and Horacio Rodríguez³

¹ Natural Language Engineering Lab,
ELiRF Research Group,
Dpto. de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia
{dbuscaldi,pross}@dsic.upv.es
² SINAI Research Group,
Computer Science Department,
Universidad de Jaén
{jmperea, laurena}@ujaen.es
³ TALP Research Center,
Universitat Politècnica de Catalunya
{dferrés,horacio}@lsi.upc.edu

Abstract. In this paper we discuss the integration of different GIR systems by means of a fuzzy Borda method for result fusion. Two of the systems, the one by the Universidad Politécnica de Valencia and the one of the Universidad of Jaén participated to the GeoCLEF task under the name TextMess. The proposed result fusion method takes as input the document lists returned by the different systems and returns a document list where the documents are ranked according to the fuzzy Borda voting scheme. The obtained results show that the fusion method allows to improve the results of the component systems, although the fusion is not optimal, because it is effective only if the components return a similar set of relevant documents.

1 Introduction

Result fusion has been studied as an option for obtaining better results in Information Retrieval (IR) by taking advantage from the combination of existing methods. Many fusion methods have been proposed, such as linear combinations [1,2] and voting schemes like the Condorcet [3] and the Borda [4] schemes. Aslam and Montague [4] concluded that the Borda fusion is a simple, unsupervised method that is capable to exceed the performance of the best component system. The fuzzy Borda voting scheme is an improvement of the standard Borda voting scheme that was introduced by [5,6]. This is the first time it is used in the IR task, although it has been used in the Word Sense Disambiguation task at Semeval⁴ with good results [7].

⁴ <http://nlp.cs.swarthmore.edu/semeval>

In Sections 2, 3 and 4 we describe briefly the systems of each group. In Section 5 we describe the fuzzy Borda ranking method, in Section 6 we present the experiments carried out and the obtained results. Finally, in Section 7 we draw some conclusions.

2 The SINAI-GIR System

The SINAI-GIR system is composed of the following subsystems: the *Collection Preprocessing subsystem*, the *Query Analyzer*, the *Information Retrieval subsystem* and the *Validator*. Each query is preprocessed and analyzed by the *Query Analyzer*, identifying its geo-entities and spatial relations and making use of the Geonames gazetteer⁵. This module also applies *query reformulation* based on the query parsing sub-task, generating several independent queries which will be indexed and searched by means of the IR subsystem. On the other hand, the collection is pre-processed by the *Collection Preprocessing* module and finally the documents retrieved by the IR subsystem are filtered and re-ranked by means of the *Validator* subsystem.

The features of each subsystem are:

- *Collection Preprocessing Subsystem*. During the collection preprocessing, two indexes are generated (*locations* and *keywords* indexes). We apply the Porter *stemmer*, the Brill POS tagger and a the LingPipe⁶ Named Entity Recognizer (NER). We also discard the English *stop-words*.
- *Query Analyzer*. It is responsible for the preprocessing of English queries as well as the generation of different query reformulations.
- *Information Retrieval Subsystem*. As IR index-search engine we have used Lemur⁷.
- *Validator*. The aim of this subsystem is to filter the lists of documents recovered by the IR subsystem, establishing which of them are valid, depending on the locations and the *geo-relations* detected in the query. Another important function is to establish the final ranking of documents, based on manual rules and predefined weights.

3 The NLEL GeoWorSE System

The system is built around the Lucene⁸ open source search engine, version 2.1. The Stanford NER system based on Conditional Random Fields [8] is used for Named Entity Recognition and classification. The toponym disambiguator is based on the method presented in [9].

During the indexing phase, the documents are examined in order to find location names (*toponyms*) by means of the Stanford NER system. When a

⁵ <http://www.geonames.org>

⁶ <http://alias-i.com/lingpipe>

⁷ <http://www.lemurproject.org>

⁸ <http://lucene.apache.org/>

toponym is found, the disambiguator determines the correct reference for the toponym. Then, the toponym coordinates are added to the *geo* index, and the toponym is stored in the *wn* index together with its holonyms and synonyms. All document terms are stored in the *text* index.

The search phase starts with the search of the topic keywords in the *text* index. The toponyms extracted by the Stanford NER are searched for in the *wn* index with a weight 0.25 with respect to the content terms. The result of the search is a list of documents ranked using the Lucene's weighting scheme. At the same time, the toponyms are used to define a geographical constraint that is used to re-rank the document list. There are two types of geographical constraints:

- a *distance* constraint, corresponding to a point in the map: documents that contain locations closer to this point will be ranked higher;
- an *area* constraint, corresponding to a polygon in the map: documents that contain locations included in the polygon will be ranked higher.

Finally, the documents retrieved by Lucene are re-ranked depending on the geographical constraints.

4 The TALP GeoIR system

The TALPGeoIR system [10] has five phases performed sequentially: collection processing and indexing, linguistic and geographical analysis of the topics, textual IR with Terrier, Geographical Retrieval with Geographical Knowledge Bases (GKBs), and geographical document re-ranking.

The collection is processed and indexed in two different indexes: a geographical index with geographical information extracted from the documents and enriched with the aid of GKBs and a textual index with the lemmatized content of the documents.

The linguistic analysis uses the following Natural Language Processing tools: *TnT*, a statistical POS tagger, the *WordNet lemmatizer 2.0*, and a *Maximum Entropy-based NERC* system, trained with the CONLL-2003 shared task English data set.

The retrieval system is a textual IR system based on Terrier [11]. Terrier configuration includes a TF-IDF schema, lemmatized query topics, Porter Stemmer, and Relevance Feedback using 10 top documents and 40 top terms.

The Geographical Retrieval uses geographical terms and/or geographical feature types appearing in the topics to retrieve documents from the geographical index. The geographical search allows to retrieve documents with geographical terms that are included in the sub-ontological path of the query terms (e.g. documents containing *Alaska* are retrieved from a query *United States*).

Finally, a geographical re-ranking is performed using the set of documents retrieved by Terrier. From this set of documents those that have been also retrieved in the Geographical Retrieval set are re-ranked giving them more weight than the other ones.

5 Fuzzy Borda Fusion

5.1 Fuzzy Borda count

In the classical (discrete) Borda count, each expert gives a mark to each alternative. The mark is given by the number of alternatives worse than it. The fuzzy variant [5,6] allows the experts to show numerically how much alternatives are preferred over others, expressing their preference intensities from 0 to 1.

Let R^1, R^2, \dots, R^m be the fuzzy preference relations of m experts over n alternatives x_1, x_2, \dots, x_n . Each expert k expresses its preferences by means of a matrix of preference intensities:

$$\begin{pmatrix} r_{11}^k & r_{12}^k & \cdots & r_{1n}^k \\ r_{21}^k & r_{22}^k & \cdots & r_{2n}^k \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1}^k & r_{n2}^k & \cdots & r_{nn}^k \end{pmatrix}$$

where each $r_{ij}^k = \mu_{R^k}(x_i, x_j)$, with $\mu_{R^k} : X \times X \rightarrow [0, 1]$ is the membership function of R^k . The number $r_{ij}^k \in [0, 1]$ is considered as the degree of confidence with which the expert k prefers x_i over x_j . The final value assigned by the expert k to each alternative x_i is the sum by row of the entries greater than 0.5 in the preference matrix, or, formally:

$$r_k(x_i) = \sum_{j=1, r_{ij}^k > 0.5}^n r_{ij}^k \quad (1)$$

The threshold 0.5 ensure the relation R^k to be an ordinary preference relation.

The fuzzy Borda count for an alternative x_i is obtained as the sum of the values assigned by each expert to that alternative: $\mathbf{r}(x_i) = \sum_{k=1}^m r_k(x_i)$.

5.2 Application of Fuzzy Borda count to Result Merging

In our approach each system is an expert: therefore, for m systems, there are m preference matrices. The size of these matrices is variable: the reason is that the document list is not the same for all the systems. The size of a preference matrix is $N_t \times N_t$, where N_t is the number of unique documents retrieved by the systems (i.e. the number of documents that appear at least in one of the lists returned by the systems) for topic t .

Each system may rank the documents using weights that are not in the same range of the other ones. Therefore, the output weights w_1, w_2, \dots, w_n of each expert k are transformed to fuzzy confidence values by means of the following transformation:

$$r_{ij}^k = \frac{w_i}{w_i + w_j} \quad (2)$$

This transformation ensures that the preference values are in the range $[0, 1]$. In order to adapt the fuzzy Borda count to the merging of the results of IR

systems, we had to determine the preference values in all the cases where one of the systems does not retrieve a document that has been retrieved by another one. The preference values of these documents were set to 0.5, corresponding to the idea that the expert is presented with an option on which it cannot express a preference.

6 Experiments and Results

In Tables 1 and 2 we show the detail of each run in terms of the component systems and the topic fields used. “Official” runs (i.e., the ones submitted to GeoCLEF) are labeled with TMESS02-08 and TMESS07A.

Table 1. Description of the runs of each system.

run ID	description
NLEL	
NLEL0802	base system (only text index, no wordnet, no map filtering)
NLEL0803	2007 system (no map filtering)
NLEL0804	base system, title and description only
NLEL0505	2008 system, all indices and map filtering enabled
NLEL01	complete 2008 system, title and description
SINAI	
EXP1	base system, title and description only
EXP2	base system, all fields
EXP4	filtering system, title and description only
EXP5	filtering system (rule-based)
TALP	
TALP01	system without GeoKB, title and description only
TALP02	complete system, including GeoKB, title and description

In order to evaluate the contribution of each system to the final result, we calculated the overlap rate O of the documents retrieved by the systems: $O = \frac{|D_1 \cap \dots \cap D_m|}{|D_1 \cup \dots \cup D_m|}$, where m is the number of systems that have been combined together and $D_i, 0 < i \leq m$ is the set of documents retrieved by the i -th system. The obtained value measures how different are the sets of documents retrieved by each system.

The R -overlap and N -overlap coefficients introduced by [12] are used to calculate the degree of overlap of relevant and non-relevant documents in the results of different systems. R -overlap is defined as $R_{overlap} = \frac{m \cdot |R_1| \dots |R_m|}{|R_1| + \dots + |R_m|}$, where $R_i, 0 < i \leq m$ is the set of relevant documents retrieved by the system i . N -overlap is calculated in the same way, where each R_i has been substituted by N_i , the set of the *non*-relevant documents retrieved by the system i .

In Table 3 we show the Mean Average Precision (MAP) obtained for each run and its composing runs, together with the average MAP calculated over the composing runs.

Table 2. Details of the composition of all the evaluated runs.

run ID	fields	NLEL run ID	SINAI run ID	TALP run ID
Officially evaluated runs				
TMESS02	TDN	NLEL0802	EXP2	
TMESS03	TDN	NLEL0802	EXP5	
TMESS05	TDN	NLEL0803	EXP2	
TMESS06	TDN	NLEL0803	EXP5	
TMESS07A	TD	NLEL0804	EXP1	
TMESS08	TDN	NLEL0505	EXP5	
Non-official runs				
TMESS10	TD		EXP1	TALP01
TMESS11	TD	NLEL01	EXP1	
TMESS12	TD	NLEL01		TALP01
TMESS13	TD	NLEL0804		TALP01
TMESS14	TD	NLEL0804	EXP1	TALP01
TMESS15	TD	NLEL01	EXP1	TALP01

Table 3. Results obtained for the various system combinations.

run ID	MAP	MAP_{NLEL}	MAP_{SINAI}	MAP_{TALP}	avg. MAP
TMESS02	0.227	0.201	0.226		0.213
TMESS03	0.219	0.201	0.212		0.206
TMESS05	0.235	0.216	0.226		0.221
TMESS06	0.226	0.216	0.212		0.214
TMESS07A	0.286	0.256	0.284		0.270
TMESS08	0.216	0.203	0.212		0.207
TMESS10	0.289		0.284	0.280	0.282
TMESS11	0.285	0.254		0.280	0.267
TMESS12	0.287	0.254	0.284		0.269
TMESS13	0.271	0.256		0.280	0.268
TMESS14	0.282	0.256	0.284	0.280	0.273
TMESS15	0.289	0.254	0.284	0.280	0.273

The obtained results show that the fuzzy Borda merging method always allows to improve the average of the results of the components, and only in two cases it cannot improve the best component result (TMESS13 and TMESS14). The results in Table 4 show that the best results are obtained if the systems returns a similar set of relevant documents (TMESS10 and TMESS12). In order to better understand this result, we calculated the results that would have been obtained by calculating the fusion over different configurations of each group's system. These results are shown in Table 5.

The fuzzy Borda method allowed also in the case of the fusion of two configurations of the same system to improve the results of the component runs. O , $R_{overlap}$ and $N_{overlap}$ values for same-group fusions are well above the O values obtained in the case of different systems (more than 0.73 with respect to 0.31 – 0.47). However, the obtained results show that the method is not able to

Table 4. O , $R_{overlap}$, $N_{overlap}$ coefficients, difference from the best system (*diff. best*) and difference from the average of the systems (*diff. avg.*) for all runs.

run ID	MAP	diff. best	diff. avg.	O	$R_{overlap}$	$N_{overlap}$
TMESS01	0.226	0.001	0.013	0.315	0.698	0.459
TMESS02	0.227	0.001	0.014	0.346	0.692	0.496
TMESS03	0.219	0.007	0.013	0.317	0.693	0.465
TMESS05	0.235	0.009	0.014	0.358	0.692	0.508
TMESS06	0.226	0.010	0.012	0.334	0.693	0.484
TMESS07A	0.286	0.002	0.016	0.356	0.775	0.563
TMESS08	0.216	0.004	0.013	0.326	0.690	0.475
TMESS10	0.289	0.005	0.007	0.485	0.854	0.625
TMESS11	0.285	0.005	0.018	0.475	0.796	0.626
TMESS12	0.287	0.003	0.018	0.356	0.822	0.356
TMESS13	0.271	-0.009	0.003	0.475	0.796	0.626
TMESS14	0.282	-0.002	0.009	0.284	0.751	0.429
TMESS15	0.289	0.005	0.016	0.277	0.790	0.429

Table 5. Results obtained with the fusion of systems from the same participant. M_1 : MAP of the system in the first configuration, M_2 : MAP of the system in the second configuration.

run ID	MAP	M_1	M_2	O	$R_{overlap}$	$N_{overlap}$
EXP1+EXP4	0.289	0.284	0.275	0.792	0.904	0.852
NLEL0804+NLEL01	0.261	0.254	0.256	0.736	0.850	0.828
TALP01+TALP02	0.283	0.280	0.272	0.792	0.904	0.852

combine in an optimal way the systems that returns different sets of relevant documents. This is due to the fact that a relevant document that is retrieved by a system and not by another one has a 0.5 weight in the preference matrix, making that its ranking will be worse than a non-relevant document retrieved by both systems.

7 Conclusions and Further Work

We combined different systems by means of the fuzzy Borda voting scheme. The implemented method allowed to improve in most cases the results of the combined systems, although the improvement was limited. The best results with this method were obtained when the systems returned a similar set of relevant documents, which means that the method needs to be improved in order to better combine sets of different relevant results. This could be done by assigning to the unknown documents a weight different from 0.5, calculating the similarity of these documents with the ones that have been retrieved by the system. This will be the focus of future research efforts.

Acknowledgements

We would like to thank the TextMESS TIN2006-15265-C06 research project and its subprojects TIN2006-15265-C06-04 MiDeS, TIN2006-15265-C06-04-05 and TIN2006-15265-C06-03 TIMOM for partially supporting this work. Daniel Ferrés is supported by a UPC-Recerca grant.

References

1. Bartell, B.T., Cottrell, G.W., Belew, R.K.: Automatic combination of multiple ranked retrieval systems. In: SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, Springer-Verlag New York, Inc. (1994) pp. 173–181
2. Vogt, C.C., Cottrell, G.W.: Fusion via a linear combination of scores. *Information Retrieval* **1**(3) (1999) pp. 151–173
3. Montague, M., Aslam, J.A.: Condorcet fusion for improved retrieval. In: CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management, New York, NY, USA, ACM (2002) pp. 538–548
4. Aslam, J.A., Montague, M.: Models for metasearch. In: SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2001) pp. 276–284
5. Nurmi, H.: Resolving Group Choice Paradoxes Using Probabilistic and Fuzzy Concepts. *Group Decision and Negotiation* **10**(2) (2001) pp. 177–199
6. García Lapresta, J., Martínez Panero, M.: Borda Count Versus Approval Voting: A Fuzzy Approach. *Public Choice* **112**(1-2) (2002) pp. 167–184
7. Buscaldi, D., Rosso, P.: Upv-wsd : Combining different wsd methods by means of fuzzy borda voting. In: Fourth International Workshop on Semantic Evaluations (SemEval-2007), ACL (2007) pp. 434–437
8. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), U. of Michigan - Ann Arbor, ACL (2005) pp. 363–370
9. Buscaldi, D., Rosso, P.: A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems* **22**(3) (2008) pp. 301–313
10. Ferrés, D., Rodríguez, H.: TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Approach with Terrier. In: Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Springer, Budapest, Hungary (2008) pp. 830–833
11. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). (2006)
12. Lee, J.H.: Analyses of multiple evidence combination. In: SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1997) pp. 267–276