Using Genetic Programming for Information Retrieval: Local and Global Query Expansion

Ronan Cummins
Dept. of Information Technology
National University of Ireland
Galway, Ireland
ronan.cummins@nuigalway.ie

Colm O'Riordan
Dept. of Information Technology
National University of Ireland
Galway, Ireland
colmor@it.nuigalway.ie

ABSTRACT

This poster presents results for two approaches using Genetic Programming (GP) to overcome the problem of term mismatch in Information Retrieval (IR). We use automatic query expansion techniques which add terms to a user's initial query in the hope that these words better describe the information need and ultimately return more relevant documents to the user.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval, Retrieval models, Query formulation. I.2.2 Artificial Intelligence: Automatic Programming

General Terms: Algorithms, Experimentation.

Keywords: Genetic Programming, Information Retrieval, Query-Expansion.

1. INTRODUCTION

In IR many relevant documents may never get returned by an IR system simply because the vocabulary of the author and that of the searcher are different. For example, when searching for information about fixing a "leaking tap", many relevant documents may be written by American authors who use the word "faucet" instead of "tap". This would typically eliminate a whole set of documents. Automatic query expansion deals with automatically adding extra terms to the query based on some heuristics and then querying the IR system again with the newly expanded query. Traditionally, this type of expansion can be categorised into local and global query expansion.

Local query expansion uses terms from the top documents of an initial retrieval run to add to the original query. Characteristics of the terms in these top few documents are used to select and weight the terms to be added to the query. Details of the GP approach to this problem can be found in [2].

Global query expansion uses terms from the corpus as a whole to add to the original query. This model of expansion assumes that terms that co-occur in many documents are semantically related. Thus, by analysing the entire collection of documents for term co-occurrences, an automatic domain-specific thesaurus can be constructed. Details of the GP approach to this problem can be found in [1].

Copyright is held by the author/owner(s).
GECCO'07, July 7–11, 2007, London, England, United Kingdom.
ACM 978-1-59593-697-4/07/0007.

2. EXPERIMENTS

We evolved term selection schemes for both approaches on one training collection (Medline) and tested them against standard benchmarks on unseen test collections. Table 1 shows a summary of results for both approaches. Mean average precision (MAP) is the measure of performance (fitness) on our test collections. The column labeled $BM25$ is the performance of the unexpanded query. LB and GB are the benchmarks for the local and global expansion approaches respectively. LGP and GGP are the best evolved solutions found on the training set for the local and global expansion approaches respectively for several runs of the GP. Overall we can see that local expansion seems to be better as it leads to an increase in performance over the original query on most collections. Our evolved local expansion scheme is comparable with the benchmark on unseen collections and surpasses it on the training collection. The global expansion approach is less successful as the performance only increases substantially on the training collection (underlined).

Table 1: %MAP for original and expanded queries

Test Collection	Queries	$BM25$	LB	LGP	GB	GGP
Medline	30	53.43	60.78	64.20	56.03	65.63
CISI	112	23.08	24.41	24.93	20.74	23.56
Cranfield	225	42.23	43.90	43.38	39.26	38.53
NPL	93	28.75	28.62	28.77	25.22	25.84

3. CONCLUSIONS

We have outlined two approaches that attempt to overcome the problem of term-mismatch in IR. Genetic programming has been used to develop formulae which automatically select terms for use in automatic query expansion techniques. The resultant scheme have been shown to outperform current benchmarks in some cases.

4. REFERENCES

[1] Ronan Cummins and Colm O'Riordan. Evolving co-occurrence based query expansion schemes in information retrieval using genetic programming. In Norman Creaney, editor, *AICS 2005*, pages 137–146, University of Ulster, 7-9 September 2005.

[2] Ronan Cummins and Colm O'Riordan. Evolving term-selection schemes for pseudo-relevance feedback in information retrieval. Technical Report NUIG-IT-201205, National University of Ireland, Galway, Ireland, 2005.