CRL English Routing System for TREC-5

Jim Cowie and Zhiqiang Guan
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
jcowie@crl.nmsu.edu, zguan@crl.nmsu.edu

Overview

The routing system described here was initially developed as a final filter to rank documents being produced by multiple queries given to a Boolean retrieval system. This system was still being developed by the time the TREC-5 deadline arrived and it was decided to use the filtering component on its own for the routing task. The router was trained using the relevance judgements from previous evaluations. It is based on a theoretical approach for discriminating between two different types of documents developed by Guthrie and Walker [1]. The generalizations made to extend this approach to multiple document types are not, however, particularly well founded. The system is a true router, each document is handled individually and matched against a profile for each topic to decide if it should be rejected or included. The results for the method are poor. The most likely cause for this is that any word in a profile can count multiple times towards success or failure. The method is similar in many ways to a vector based retrieval system, with a vector for each topic being matched against a word vector for the document.

Method

Word and two word frequency lists are produced for each relevant document set and for "general" text (all documents). These are sorted in descending frequency and words (or bi-grams) chosen which occur in the first n words in the list for relevant documents and after the first m words in the list for non-relevant documents. These two values were finally set at 500 and 1000. A list of words is chosen in the same way to represent non-relevant documents. Thus we have a list of words occurring frequently in relevant documents and infrequently in relevant documents and a second list of words occurring frequently in non-relevant documents and infrequently in relevant documents. We know the frequencies of each word in each list for both types of documents and the total number of words in both types. We can thus calculate the frequency of occurrence of each type of word; relevant, non-relevant, and words which are not considered significant, for both document types (relevant and non-relevant).

These are used to estimate if a documents word profile belongs to relevant, or non-relevant by calculating

$$(n1 \log p11 + n2 \log p12 + n3 \log p13)$$

to estimate if a document is relevant, where n1 is the number of "relevant" words and p11 their relative frequency in the relevant document training data, n2 is the number of "non-relevant" words and p12 their relative frequency in the relevant document training data and n3 is the number of other words and p13 is their relative frequency in the relevant document training data. A similar measure is made to estimate if the document is non-relevant. If the difference between the two measures is greater than zero the document is passed to a second ranking filter based on bi-grams. This second does not reject documents, but gives each a score by which they are ordered. This order was based on various trials which seemed to indicate that a word based measure gave better recall, and the bi-gram based method gave better precision.

Word Lists

The word lists generated are about 100 items long and many of the words are obviously related to the topic. Temporary occurrences (e.g. a particular incident occurring in a certain location) also show up in the lists and some method is needed to remove these, as they are in fact topic independent. Alternatively each word should only be allowed to count once in a document and the relative frequencies used should be number of documents containing a word divided by the number of documents. We intend to experiment further with this technique.

Example Lists

Partial lists are given below for the poaching (77) and smoking cessation (173) topics.

Topic - 77 - Top relevant words

573 IVORY
426 WILDLIFE,
254 POACHERS,
253 FISHING
245 ELEPHANT
237 POACHING
232 ELEPHANTS
222 FISH
199 SPECIES
199 HUNTING
196 ANIMALS
184 AFRICAN
175 ENDANGERED
166 BAN
164 ILLEGAL
146 AFRICA
144 BEAR
126 PARK
124 KENYA
123 POPULATION
115 RHINO
112 COAST
111 TUSKS
105 ANIMAL
99 CONSERVATION
97 EST

Topic - 77 - Top relevant bi-grams

254 U_S
100 ENDANGERED_SPECIES
91 NR_EST
74 NR_EDT
67 EST_R
67 EDT_R
56 NATIONAL_PARK
54 FISH_WILDLIFE
49 AFRICAN_ELEPHANT
48 WILDLIFE_SERVICE
45 IVORY_TRADE
43 ELEPHANT_POPULATION
39 TRADE_ENDANGERED
38 NORTH_KOREA
37 NORTH_KOREAN
35 NATIONAL_PARKS
33 WORLD_WILDLIFE
32 WILDLIFE_FUND
29 S_FISH
29 ELEPHANT_IVORY
28 GALL_BLADDERS
27 IVORY_IMPORTS

Topic - 173- Top relevant words

2779 SMOKING
1624 TOBACCO
1115 BAN
635 SMOKE
635 ADVERTISING
624 FLIGHTS
613 CIGARETTE
568 SMOKERS
347 AIRLINE
341 ADS
331 CIGARETTES
309 AD
289 AIRLINES
256 FLIGHT
222 DOMESTIC
210 PHILIP
184 RESTRICTIONS
181 MORRIS
180 RJR
179 TRANSPORTATION
172 CANCER
165 MEASURE
160 PASSENGERS
155 SPENDING
155 LEGISLATION
150 NORTHWEST
149 INSTITUTE
147 BANNED
146 SAATCHI

Topic - 173 - Top relevant bi-grams

1461 U_S
336 SMOKING_BAN
262 ANTI_SMOKING
174 TOBACCO_PRODUCTS
173 PHILIP_MORRIS
170 BAN_SMOKING
166 TWO_HOURS
157 TOBACCO_INDUSTRY
120 HOURS_LESS
113 TOBACCO_ADVERTISING
110 DOMESTIC_FLIGHTS
107 NR_EDT
102 TOBACCO_COMPANIES
96 NR_EST
95 PRODUCTS_TOB
95 NON_SMOKERS
90 AIRLINE_FLIGHTS
88 SURGEON_GENERAL
85 SMOKE_FREE
80 SMOKING_DOMESTIC
79 NON_SMOKING
78 SMOKING_AREAS
75 FLIGHTS_TWO

Performance

The combined recall-precision measure over all topics was 0.20. This is a poor level of performance and indicates serious flaws in the method.

Acknowledgement

The work described here was funded by the DoD contract MDA904-91-C-6153.

Reference

[1] Guthrie, L, Walker, W,. and Guthrie J., Document Classification by Machine, MCC-92-235, Computing Research Laboratory, New Mexico State University, 1992.