DCU at NTCIR-8 GeoTime

Zhengwei Qiu, Cathal Gurrin, Aiden R. Doherty, Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland
{zqui, cgurrin, adoherty, asmeaton}@computing.dcu.ie

ABSTRACT

In this report, we describe the experiments carried out by Dublin City University for NTCIR GeoTime 2009-10. In all we submitted five runs, which evaluated the benefit of including clustered location information when compared to a standard word-term text IR ranking. The baseline technique was the Lucene default and we developed three different algorithms to re-rank the results based on location occurrance. In conclusion we found that the inclusion of location information to re-rank documents only offered a minor improvement over the baseline. Our analysis leads us believe that larger gains can be made through including location information at the indexing and initial querying stage, and then refining the final ranked list by standard IR techniques.

1. INTRODUCTION

In our research for NTCIR 2009-10, the GeoTime task [5], we were interested in examining if the inclusion of location specific analysis aided in the processing of geographic and temporally restricted queries. The dataset employed for NTCIR GeoTime 2009-10 consisted of a collection of 315,417 news items from The New York Times 2002-2005 and 25 queries. Each of the 25 queries contains three components; the ID, the description and the narrative, all of which are in both English and Japanese. For our work, we only utilize the English version. After initial experimentation, we found many challenges in extracting accurate temporal information only based on regular expressions and the temporal information that we did extract,we found to be less useful than we expected in our initial experimentation. Therefore, the focus of this work was in utilizing the location information to improve the quality of retrieval over a content-only baseline. However, as will be described, we do retain the temporal information when we preprocess the queries. Because there is no obvious spacial location aspect in the topics (e.g. information on tourist sites within 5KMs of Tokyo), we could not use the gazetteer to improve recall of potentially relevant documents,as we could do to other types of Geographical Information Retrieval (GIR) queries. However, we could adopt different ranking algorithms for locations to re-rank the original top 1,000 relevant documents as originally ranked by a search engine. The different ranking algorithms will be described in the later section on re-ranking. The remainder of this paper is as follows: in section 3, we describe the process of creating index. In section 4, we will discuss query-preprocessing. In section 5, we describe Topic Post-Processing. In section 6, we describe the experiments into reranking the result of the baseline text content only run. In section 7, we will discuss the results and finally we present our conclusions and future plans.

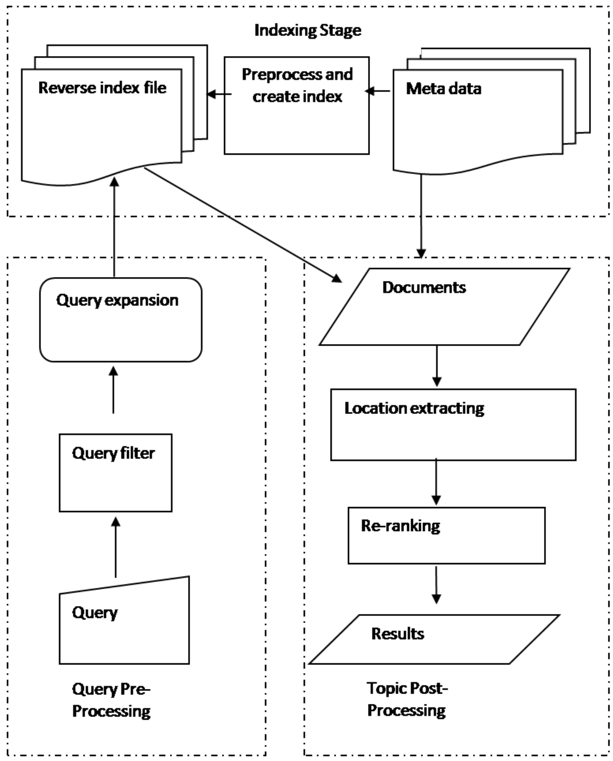

Figure 1: system architecture

2. SYSTEM OVERVIEW

There are three main parts in our system. They are Indexing, Query Preprocessing and Topic Post-Processing. As shown in Figure 1, the main function of the Indexing Stage is to deal with data files such as the preprocessing of xml files, creating the index and also reformulating the query. The Query Preprocessing is to filter and to expand the queries for one run which was concerned with query expansion. The Topic Post-Processing is to re-rank the results which are returned from the search engine using a number of different, but related techniques. The operation of the system components will be described in the next sections.

3. INDEXING STAGE

```
<DOC id="NYT_ENG_20020101.0001" type="story" >
<HEADLINE>
JANICE FARRAR THADDEUS, 68, LITERARY SCHOLAR
</HEADLINE>
<DATELINE>
 (BC-OBIT-THADDEUS-NYT)
</DATELINE>
<TEXT>
<P>
Janice Farrar Thaddeus, a scholar, poet, editor and former
Harvard lecturer in English, died Dec. 23 in Cambridge, Mass. She
was 68.
</P>
<P>
The cause was a stroke, said her husband, Patrick Thaddeus.
</P>
<P>
In ``When Women Look at Men'' (Harper Row), the feminist
anthology Dr. Thaddeus edited in 1963 with John A. Kouwenhoven, she
and her co-editor provided an answer to the question why women do
not achieve as much as men in the professions: ``They have no
wives.''
</P>
<P>
Thaddeus also wrote scholarly articles on Jonathan Swift,
Richard Wright and Anthony Burgess.
</P>
<P>
She wrote a biography, ``Frances Burney: A Literary Life'' (St.
Martin's, 2000), about a writer who had been considered a minor
novelist of manners even though she was admired by Samuel Johnson,
Edmund Burke, Jane Austen and Edward Gibbon.
</P>
```

Figure 2: Original document

We will first discuss document and query preprocessing before moving onto describing topic post-processing and location experiments.

3.1 Preprocessing

From the original documents, we can extract some useful infomation, such as the document id, the title of the news and also the time of publishing from the XML format files. This information could not be used directly for our system, and was therefore converted into common text for later indexing. Fortunately, there are many tools and libraries which can be employed to extract the information from XML and remove the tags from the documents. After preprocessing, the cleaned documents amounted to about 2.1G. The entire process is described below:

1. A sample document prior to preprocessing is shown in Figure 2. Every document has an id, title and content.

2. In the process, all the contents between “<” and “>” are removed. And also some are translated to their corresponding word. Such as “&” is converted to “&”.

3. The same sample document after preprocessing is shown in Figure 3.

3.2 Creating index

When we get the cleaned documents from the Preprocessing stage (section 3), we add the document into the inverted indexof the search engine, For this we use the lucene search engine and emply the default vector space model [6].

4. QUERY PREPROCESSING

As known, not all words in the query are equally useful for generating ranked output in information retrieval. For this work, the most useful words in the sentences, we believe to

```
JANICE FARRAR THADDEUS, 68, LITERARY SCHOLAR

(BC-OBIT-THADDEUS-NYT)

Janice Farrar Thaddeus, a scholar, poet, editor and former
Harvard lecturer in English, died Dec. 23 in Cambridge, Mass. She
was 68.

The cause was a stroke, said her husband, Patrick Thaddeus.

In ``When Women Look at Men'' (Harper & Row), the feminist
anthology Dr. Thaddeus edited in 1963 with John A. Kouwenhoven, she
and her co-editor provided an answer to the question why women do
not achieve as much as men in the professions: ``They have no
wives.''

Thaddeus also wrote scholarly articles on Jonathan Swift,
Richard Wright and Anthony Burgess.

She wrote a biography, ``Frances Burney: A Literary Life'' (St.
Martin's, 2000), about a writer who had been considered a minor
novelist of manners even though she was admired by Samuel Johnson,
Edmund Burke, Jane Austen and Edward Gibbon.
```

Figure 3: Cleaned document

be the nouns, verbs and also adjectives. Therefore, in this step we employ the Stanford parser to analyze every query [3]. For example, when we process the query of *“When and where did a volcano erupt in Africa during 2002?”*, we get output of the parser is *“When/WRB and/CC where/WRB did/VBD a/DT volcano/NN erupt/VB in/IN Africa/NNP during/IN 2002/CD ?/.”* In this example “/NNP” means proper noun, “/NN” means noun, “/VB” means verb, “/WRB” means Wh,“/CC” means conjunction,“/VBD” means past tense of verb, “/DT” means article ,and “/CD” means number. For the work presented here, the most useful words are *“volcano, erupt, Africa,2002”*. However, even these words can not be considered to be equally useful. For the words of “volcano, erupt”, because of spelling variations and synonyms degrades, they can be replaced easily by other words. Hence some relevant documents may not be retrieved or are ranked very low in the ranked list when the author uses different words to describe an event or represent an information need. On the other hand, the word “Africa” like other location, personal and organization names, has special meaning without synonyms. Therefore it should remain unmodified in the query. “2002” should equally be fixed. We get the new queries by applying the filter on the original. For run 3, we apply a query expansion to new queries based on WordNet [4]. Query expansion may improve recall, but also increase noise in the result set because of the added synonyms may move the query off-topic somewhat. In some situations, incorporating such a query expansion step will negativly impact on precision in a major way, therefore we do not apply the query expansion for every run [7]; in fact only one run (run 3) employs query expansion.

5. TOPIC POST-PROCESSING

Similar to the creating index stage (see Section 3), the step of topic retrieval is implemented mainly through the Lucene search engine library [2]. After the queries are preprocessed, the new queries are sent to search engine for run 1. The top 1,000 relevant documents will be returned. This ranking will be seen as the baseline ranking and all of the subsequent rankings based on utilising location data will depend on this baseline. Another set of 1,000 document is generated for run 3 as there is no re-ranking required for run 3.
From the baseline's names, our process applies Lingpipe [1] with its Named Entity Recognition model to extract all the

locations for every document. In this process, we remove the repeated location names, but record the frequency of each location. The list of location and their frequencies will be used for re-ranking the documents. The process is shown below where the original document is:

... For the first time since it was granted the power in the 1850s, the Mexican Senate on Tuesday held up the foreign travel plans of the head of state, prohibiting President Vicente Fox from visiting the United States and Canada next week.
Political analysts said the action was part of increasingly heated disputes over foreign policy matters between the Fox administration and the Congress. Those include the extent of Fox's travels abroad and strains in relations with Cuba as Mexico strengthens its ties to the United States.
For months, members of Congress have expressed concern about Fox's trips, which have included visits to Asia, Europe and South America, accusing him of spending more time seeking foreign investment and international stature than working on domestic issues...

The output of Lingpipe is :

<output>
<s i="0">
<ENAMEX ID="0" TYPE="ORGANIZATION">
MEXICO BARS ITS PRESIDENT FROM TRIPS NEXT WEEK MEXICO CITY
</ENAMEX>
(
<ENAMEX ID="1" TYPE="ORGANIZATION">
BC-MEXICO-PRESIDENT-NYT) For
</ENAMEX>
the first time since it was granted the power in the 1850s, the Mexican Senate on Tuesday held up the foreign travel plans of the head of state, prohibiting President
<ENAMEX ID="2" TYPE="PERSON">Vicente Fox
</ENAMEX>
from visiting the
<ENAMEX ID="3" TYPE="LOCATION">United States
</ENAMEX>and
<ENAMEX ID="4" TYPE="LOCATION">Canada
</ENAMEX>
next week.</s>
- <s i="1">
Political analysts said the action was part of increasingly heated disputes over foreign policy matters between the
<ENAMEX ID="2" TYPE="ORGANIZATION">Fox
</ENAMEX>
administration and the
<ENAMEX ID="5" TYPE="ORGANIZATION">Congress

</ENAMEX>.</s>
- <s i="2">
Those include the extent of
<ENAMEX ID="6" TYPE="ORGANIZATION">Fox's
</ENAMEX>
travels abroad and strains in relations with
<ENAMEX ID="7" TYPE="PERSON">Cuba
</ENAMEX>
as Mexico strengthens its ties to the
<ENAMEX ID="3" TYPE="LOCATION">United States

Figure 4: Location list

</ENAMEX>.</s>
</output>

In the output of the Lingpipe, *"<s i='0'>"* annotates the first sentence of the input, and *"<ENAMEX ID="0" TYPE="ORGANIZATION"> "* means the name entity's ID and type. By analysing the output, we get a location list which will be used in re-ranking section like Figure 4.

6. LOCATION EXPERIMENTS

Before re-ranking the documents, we evaluate 3 ranking technique:

1. Rank documents with at least one location higher than no location.
2. Rank documents highly if they have the most frequently occurring location.
3. Rank documents higher based on a measure of location diversity and novelty, which can bring potentially relevant new information to the user [9].

6.1 Relevant Documents with at least One Location

For the first ranking technique, we re-rank all the documents based on the original ranking (from Section 4) by moving the documents without location to the bottom of the ranked list. In Figure 5 and the subsequent figures, the word begins with "L" indicates location; the "NL" means the doucument without location. As can be seen, D4 and D6 (both NL) are moved (in the original rank order) to the bottom of the reranked list. The number following "D" indicates the original baseline rank of the document.

6.2 Most Relevant Document Will Come From Most Commonly Occurring Location

For the second ranking technique, we calculate a ranking of locations based on their frequency of occurrance in the baseline set. Thereafter we cluster all the documents by their detected location. From this we infer that the location cluster containing the most documents is the location which is most likely to be relevant to the given task. Within this cluster the highest ranked original document is now ranked first for the task, the 2^{nd} highest in this cluster is 2^{nd} most likely unto all n documents in the cluster have been regarded as the top n most likely overall to the task. In the same way, the second largest cluster of locations, then contains the documents to be ranked from positions $n+1$ to $(n+1)+o$, where o is the number of documents in the second largest cluster of locations. After obtaining all the locations' relevant documents, we put the rest of the documents (that have no identified location) in the tail of the new ranked list

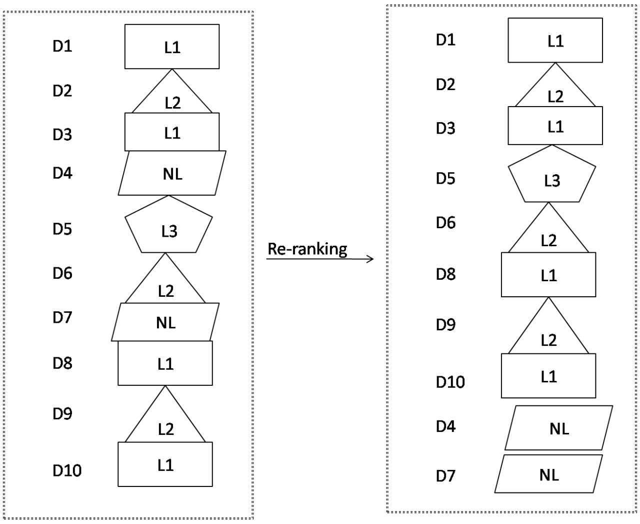

Figure 5: Re-ranking by location presence

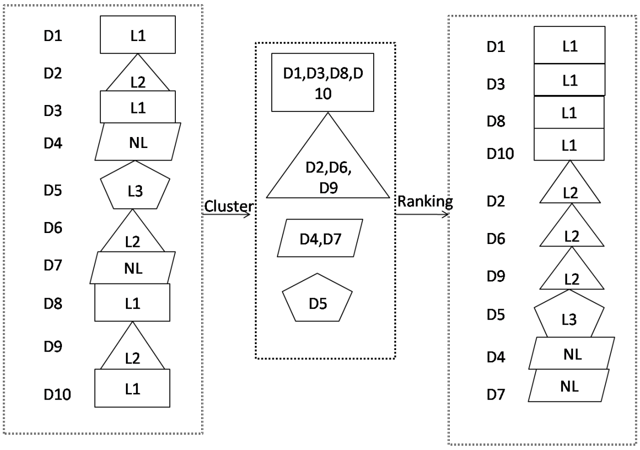

Figure 6: Re-ranking in location important order

(in baseline rank order).This process is visualised in Figure 6.

6.3 Top N Documents Containing Diversity of Locations

In this ranking technique we investigate whether highly ranking a diverse set of documents from many different locations introduces novel and potentially interesting results for the user. Again search in the original ranking and find those documents with locations. As in section 6.2, we cluster all the documents by their detected location. From this we infer that the location cluster containing the most documents is the location is most likely to be relevant to the given task. Within this cluster the highest ranked original document is now ranked first for the task, but the 2^{nd} ranked document overall from the task comes from the highest rank document in the 2^{nd} largest cluster, the 3^{rd} ranked document overall comes from the highest ranked document in the 3rd largest cluster, and so on unto all c clusters have been selected from. Then overall document ranked $c+1$ comes from the 2^{nd} ranked document in the largest cluster of locations, $c+2$ comes from the 2^{nd} ranked document in the 2^{nd} largest cluster, $c+3$ from the 2^{nd} ranked document in the 3^{rd} largest cluster, and so on. After obtaining all the locations' relevant documents, we put the rest of the documents (that have no identified location) in the tail of the new document's list. This process is visualised in Figure 7.

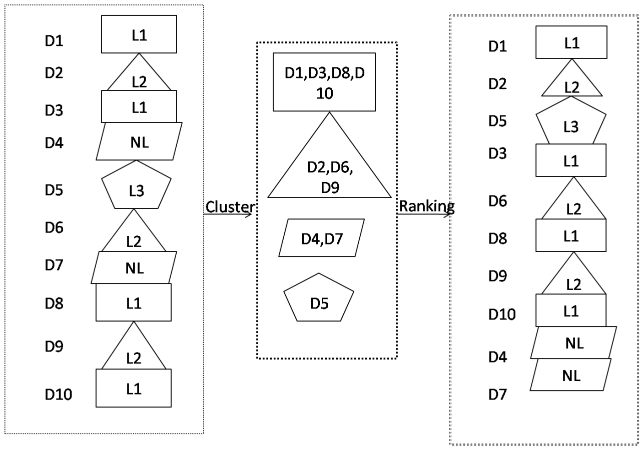

Figure 7: Re-ranking in round robin order

Table 1: Topic relevance results

runID	mean AP	mean Q	mean nDCG
DCU-EN-EN-01-D	0.3207	0.3404	0.5506
DCU-EN-EN-02-D	0.3218	0.3413	0.5513
DCU-EN-EN-03-D	0.2807	0.2991	0.5129
DCU-EN-EN-04-D	0.2491	0.2593	0.4843
DCU-EN-EN-05-D	0.241	0.2643	0.5042

7. RESULTS

We submitted 5 runs as follows:

1. DCU-EN-EN-01-D: A baseline run only using the title of query and automatically retrieved by Lucene (Section 5).

2. DCU-EN-EN-02-D: The baseline, reranked by location presence (Section 6.1).

3. DCU-EN-EN-03-D: Search engine ranking; based on WordNet query expansion (Section 5).

4. DCU-EN-EN-04-D: The baseline, reranked by the location clustering in decreasing order of location frequency (Section 6.2).

5. DCU-EN-EN-05-D: The baseline, reranked by location clustering, in decreasing order of occurrance freequency, in round-robin order (Section 6.3).

The evaluation results for the 5 submitted runs are listed in Table 1 and also visualised in Figure 8.

Our initial analysis of the results appears to show that the baseline Lucene query ranking (DCU-EN-EN-01-D) can be improved by reranking the output by those documents containing locations (DCU-EN-EN-02-D), with performance never being below that of the baseline run shown in Figure 8. However once the "non-location" documents have been filtered out, the original query-term based ranking appears

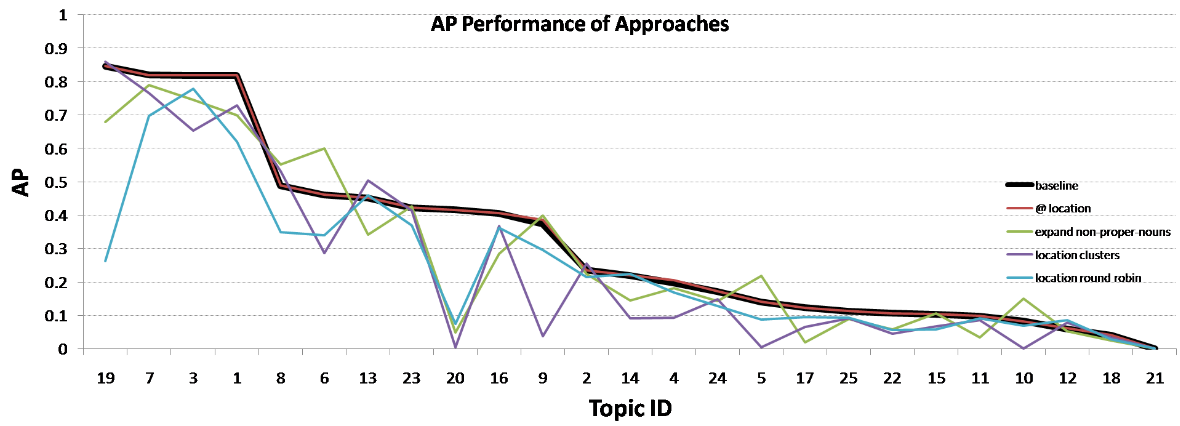

Figure 8: Re-ranking in round robin order

Table 2: Correlation between proper nouns, common nouns, and verb present in each query and final AP performance for each submitted run

	proper nouns	common nouns	verb
1. baseline	0.32904	-0.3719	0.29163
2. top n location	0.33148	-0.3761	0.29561
3. wordnet	0.38136	-0.35743	0.38616
4. location cluster	0.32015	-0.19265	0.14680
5. round robin	0.33187	-0.33893	0.45234

more suitable in ranking the results appropriately (DCU-EN-EN-02-D), rather than using locational inferences (e.g. clustering by location, etc.) (DCU-EN-EN-04-D, DCU-EN-EN-05-D), as seen in detail in Figure 8.

Considering the number of proper nouns (NNP), common nouns (NN), and verbs (VB) associated with each query, we then investigated the correlation between the number of each of these, and also the AP performance of each run across all 25 queries. The results of this are detailed in Table 2.

From Table 2 we can see that run 3 (DCU-EN-EN-03-D) is most effected with regards to number of proper nouns present in each query. However we must consider that we only expand the other words (i.e. those that aren't proper nouns) in those queries. We believe this merits further investigation in the future to better understand the effect taking place here.

Also from Table 2 we can see that run 5 (DCU-EN-EN-05-D) is sensitive with regards to the number of verbs present in a query (correlation of 0.45), with more verbs meaning higher precision scores. Again we believe that further investigation is required into understanding the meaning of this effect, which will help guide future developments.

8. CONCLUSIONS AND FUTURE WORK

In this work we have investigated whether employing geographic-based reranking of a baseline text only result set can improve performance over the baseline ranking alone in the geotime task of NTCIR 8. In all we evaluated three different reranking techniques and one query expansion technique and found that the ranking can be improved by removing the documents without locations and also query expansion can improve some topics' ranking in the Geotime task.

We believe the path forward in the future is to firstly find relevant documents with location inferentials (e.g. location entity recognition, and possibly clustering occurring in the "pre-ranking" stage) containing a higher degree of influence, and then to rank the filtered results using standard query-term ranking algorithms.

Indeed this is beginning to mirror results in the domain of finding digital lifelog information, whereby location is a powerful inferential guide to locate a relevant cluster of content, but then other approaches are needed to rank that material in a more fine-grained fashion [8].

Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1147.

9. REFERENCES

[1] Lingpipe. http://alias-i.com/lingpipe.

[2] Lucene search engine package. http://lucene.apache.org.

[3] The stanford parser. http://nlp.stanford.edu/software/lex-parser.shtml.

[4] Wordnet. http://wordnet.princeton.edu.

[5] N. K. J. M.-F. T. S. Fredric Gey, Ray Larson. Ntcir-geotime overview: Evaluating geographic and temporal search, 2010.

[6] C. S. Y. G. Salton, A. Wong. A vector space model for automatic indexing. In *Communications of the ACM*, pages 613 – 620, New York, NY, USA, 21975. ACM.

[7] I. C. J. C. Julio Gonzalo, Felisa Verdejo. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP*, 1998.

[8] V. Kalnikaité and S. Whittaker. Software or wetware?: discovering when and why people use digital prosthetic memory. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 71–80, New York, NY, USA, 2007. ACM.

[9] H. R. Varian. Economics and search. pages 1–5. SIGIR Forum, 1999.