DUTIR at TREC 2007 Blog Track

Song Rui, Tang Qin, Daming Shi, Hongfei Lin, Zhihao Yang
Department of Computer Science and Technology, Dalian University of Technology
No 2 LingGong Road Shahekou District, Dalian 116023, China.
xiaorui84315@sina.com, woaiwojia8212@163.com, damingshi@gmail.com,
{hflin, yangzh}@dlut.edu.cn

Abstract

This paper describes DUTIR at TREC 2007 Blog Track. In data preprocessing, a non English language list created from the corpus was used to remove the non English blogs, blog templates were also used to extract the post and comment; in Opinion Retrieval task, information in the meta tags were also indexed; in the polarity subtask, a method based on SVM was used and the Information Gain attribute selecting method was used to assist SVM; in Feed Distillation task, three type of feeds were analyzed according to their tag structure, information extracted from particular tags of the feeds were finally indexed.

1. Introduction

DUTIR (Information Retrieval Laboratory of Dalian University of Technology) participated in all of the tasks of the TREC 2007 Blog Track, including Opinion Retrieval Task, Polarity Subtask and Feed Distillation Task. Modules designed for different tasks are described separately below.

2. Opinion retrieval task

In Opinion Retrieval Task, the main retrieval unit is permalink. The processing was divided into three steps: Data Preprocessing, Indexing, and Query Construction.

2.1 Data Preprocessing

In this step, non English Blogs were removed, useless tags such as Script and Style were also removed, and some blog templates which have a relatively large proportion in corpus were made use of to extract the desired content accurately.

A non English list which contains about 120 non English languages was created from the corpus. Feeds and its permalinks that belong to these non English languages were removed. Table.1 shows the non English feeds extracted by the non English list. The Non English feed number is 37625 accounting for 4.99% of the total feed number (753,681).

Total Feed Number	753,681
Non English Feed Number	37625
Non English Feed Proportion	4.99%

Table.1 Non English Feeds Proportion

Tags such as Script, Style and so on are useless and contribute to a large proportion of the corpus. Some simple pattern matching techniques were used in order to remove these tags.

Blogs generated by programs follow well defined markup rules allowing the post's content to be identified [1]. So some blog templates which contribute to relatively a larger proportion of the corpus such as Blogger, WordPress and so on were manually read and summarized into particular templates. However, because of the large corpus, we were unable to construct all of the templates, only blogs with top

proportions were extracted into some templates, the remains were parsed off all of the html tags with toolkit htmlparser [2]. Table.2 below show the other results of the Data Preprocessing.

Total Permalink Number	3,215,171
Non English Permalink Number	173026
Non English Permalink Proportion	5.38%
Size of total permalinks	88.7GB
Size of total permalinks with Non English Permalinks and <DOCHDR> tags removed	81GB
Size of total permalinks with Script and Style tags removed moreover.	66GB
Size of total permalinks with whole or parts of tags removed according to some blog templates	16.7GB

Table.2 Permalink Preprocessing Result

2.2 Indexing

The preprocessed permalinks were later indexed with Indri Search Engine [3]. Besides the desired content between the tags, keywords and description parts of the Meta tags were also indexed as another two fields.

2.3 Query construction

In this step, two tasks were focused on: constructing a sentimental lexicon and extending the query.

2.3.1 Construction of sentimental lexicon

Although there were already some sentimental lexicons that have been constructed, such as Welsh to English Lexicon [4], General Inquiry [5] and SentiWordNet [6] and so on, most of the words in the lexicons don't emerge frequently in corpus because of the casual language style of blogs. So instead of using one of the lexicons, we choose to build our own lexicon which contains about 2000 sentimental words that emerge frequently in the corpus.

2.3.2 Query expansion

When the queries were being constructed, exact or partially exact match of words in the title parts were consider first and this will contribute to the topic relevance of the results. However, not all of the topics could get enough results with the exact or partially exact match, in this case, query extension including adding some key words in description and narrative fields to the query or making use of some knowledge found in knowledge base on Internet, such as Wikipedia were used. For example, there is a topic about a band and we need to find all of the blogs that expressed some opinions about this band or its members. Necessary information such as the names of the members needs to be found on Internet due to the inexistence of this information in description and narrative fields. Query feedback was also tried; however, results were disappointing. A simple method was used to find the blogs expressing opinions about a topic, which is judging by the emergence of sentimental words around topic words [7].

2.4 Results

Table.3 shows the six Runs submitted in Opinion Retrieval Task. From Table.3, the runs with opinion words added to assist in finding opinions behaved better than those without. Moreover, more retrieval fields will not bring better results, DUTRun4, DUTRun5 and DUTRun6 relatively behaved worse than DUTRun1, DUTRun2 and DUTRun3. More fields mean more noise. However, moderately using the restriction in other fields will improve the results.

RunID	Description	MAP	R-Precision	P@10
DUTRun1	Automatic and title-only	0.2890	0.3368	0.502
DUTRun2	Title +Opinion Words	0.3190	0.3671	0.6
DUTRun3	Title +Description+ Opinion Words	0.3094	0.3527	0.6060
DUTRun4	Title +Description	0.2843	0.3360	0.4820
DUTRun5	Title+ Description +Narrative	0.2279	0.3029	0.5060
DUTRun6	Title +Description +Narrative + Opinion Words	0.2959	0.3401	0.6080

Table.3 Opinion runs submitted

Table.4 shows the top 5 improvements of the best submitted compulsory automatic title-only runs over the baselines.

Group	Best Baseline	Baseline Map	Best Non-baseline	Non Baseline MAP	%Increase
UGlasgow(Ounis)	uogBOPFProx	0.2817	uogBOPFProxW	0.3264	15.87%
IndianaU (Yang)	Oqsnr1Base	0.2537	oqsnr2opt	0.2894	14.07%
UArkansas Littlerock (Bayrak)	UALR07Base	0.2554	UALR07BlogIU	0.2911	13.98%
DalianU (Yang)	DUTRun1	0.289	DUTRun2	0.319	10.38%
UWaterloo (Olga)	UWbasePhrase	0.2486	UWopinion3	0.2631	5.83%

Table.4 The top 5 improvements over the baselines for automatic title-only runs

According to Table.4, the opinion finding method in this paper seems to be helpful, improving our performance on the task by 10.38%, despite our good performing baseline.

3. Polarity subtask

In this task, a method based on Information Gain (IG) and Supporting Vector Machine (SVM) was finally used to judge the polarity of blogs. With the 2006's 50 queries and their associated relevance judgments, there was no need to find training set. One of the original experiments which were later proved to be ineffective was to train each document with the whole content. Often there were more than one topic in a document; to train the whole document seems to be confused. So sentences with topic words and their pronouns were extracted as the final training set. During the accomplishment of this task, several tests were tried including method based on sentimental lexicon, method based on machine learning and the combination of the two.

3.1 Method based on sentimental lexicon

This method is mainly judging the polarity by the distribute density of sentimental words. However, some documents which were labeled as negative contained many positive words. This made the method only based on the distribute density of sentimental words seem not much too accurate.

3.2 Method based on machine learning

A method based on SVM was used to train the corpus. Feature selection is necessary before training. Since this is a sentimental classification problem, sentimental words in the lexicon were considered as feature words first. However, the classification accuracy was not very satisfying. Feature words selected by Information Gain were later used as features to train the corpus. Moreover, combination of the two was also tried. The whole classification was divided into two steps: First, classify the training set into opinionated ones and non opinionated ones; second, classify the opinionated ones into positive, negative and the combination of the two

3.3 Results

Table.5 shows the six polarity runs submitted. Because of the correspondence between polarity runs and opinion runs, it seems that the runs behave better in opinion retrieval task also behave better in polarity task.

Run	R-Acc	A@10	A@1000
DUTRun1P	0.1603	0.2708	0.0412
DUTRun2P	0.1721	0.3080	0.0406
DUTRun3P	0.1624	0.3040	0.0400
DUTRun4P	0.1608	0.2620	0.0406
DUTRun5P	0.1356	0.2520	0.0255
DUTRun6P	0.1591	0.3020	0.0380

Table.5 Polarity runs submitted

Table 6 shows the top 5 best-scoring title-only polarity detection run for each group in terms of R-accuracy, regardless of the topic length, and sorted in decreasing order of R-accuracy.

Group	Run	R-Acc	A@10	A@1000
UIUC (Zhang)	uic75cpnm	0.2295	0.3700	0.0493
UAmsterdam (de Rijke)	uams07ipolt	0.1827	0.2640	0.0418
IndianaU (Yang)	oqsnr2optP	0.1799	0.2800	0.0401
DalianU (Yang)	DUTRun2P	0.1721	0.3080	0.0406
Zhejiangu (Qiu)	EAGLE2P	0.1510	0.2380	0.0427

Table.6 The top 5 best polarity runs for each group in terms of R-accuracy

According to Table.6, systems which are more successful at retrieving opinionated documents ahead of relevant ones, they will then have more documents for which they can make a correct classification. The five groups in this table also performed well in the opinion retrieval task.

4. Feed distillation task

In this task, feed files were preferred as retrieval units due to the need to submit feedno. Like the opinion retrieval task, processing was divided into three steps: Data Preprocessing, Indexing, and Query Construction.

4.1 Data preprocessing

Some manual work such as analyzing the possible formats of the feed files and recording the possible tags that contain desired information according different formats are necessary. Finally three types of feed formats were found according their different displaying styles: RSS, RDF, and ATOM. Moreover, some types of feed could be divided into smaller units such as Item or Entry. Whether for the whole feed or the smaller units in it, desired contents in them are often in the fixed tags such as <description>, <content>, <summary> and so on. So feed files with non English ones removed were later parsed according these tags with htmlparser iteratively. When using htmlparser, these tags need to be defined and registered in order to identify these tags and extract contents among them. The size of original feed is 38.8GB and the size of feed after non English ones and undesired contents are removed becomes 17.5GB.

4.2 Indexing

Preprocessed data were later indexed with Indri Search Engine [3]. Feeds are different from permalinks, there are often redundant feeds among different files and the contents of them are varying or not. So when run query on the index, it needs to remove the redundant feeds and make a little adjustment.

4.3 Expanding the queries

In this step, some key words in description and narrative fields were utilized to expand the queries, knowledge found in some knowledge base were also utilized. Since this task is different from opinion retrieval task which needs to find all of the opinionated blogs about a particular topic, moreover, content of each topic that belongs to this task is quite wide, news also meet requirement, so not each query was constructed with sentimental words added.

4.4 Results

Table.7 shows the top 5 best-scoring automatic title-only run from each participating group in terms of MAP, and sorted in decreasing order.

Group	Run	MAP	R-prec	b-Bref	P@10	MRR
CMU (Callan)	CMUfeedW	0.3695	0.4245	0.3861	0.5356	0.7537
UGlasgow (Ounis)	uogBDFeMNZP	0.2923	0.3654	0.3210	0.5311	0.7834
UMass (Allen)	UMaTiPCSwGR	0.2529	0.3334	0.2902	0.5111	0.8093
KobeU (Seki)	Kudsn	0.2420	0.3148	0.2714	0.4622	0.7605
DalianU (Yang)	DUTDRun1	0.2285	0.3105	0.2768	0.3711	0.5813

Table.7 Blog distillation results: the top 5 automatic title-only run with the best MAP, sorted by MAP

As shown in Table.7, the method in this paper to find feeds was not satisfying. During the manually labeling phase, this fact was already found. A lot of feeds that contently relate to particular topic are in fact do not have enough even no correspond permalinks, while our method only focused on the content of feeds, so the final result was consequently not satisfying. We are confirmed that the addition of the consideration of permalinks to the method that used in this task will perform better.

5. Conclusions

After analyzing and comparing the results, some conclusions can be drawn. When finding opinions, a sentimental lexicon which contains opinion words frequently emerged in blog corpus is necessary. Title field is the basic field, other fields are not as important as title field and should be used with caution in case of bring noise. It seems that simply considering feed as retrieval unit do not perform well in Feed Distillation task.

References

[1]Blog Mining through Opinionated Words, G. Attardi, M. Simi, Università di Pisa.

[2]HTML Parser Tool Kit: http://sourceforge.net/project/showfiles.php?group_id=24399.

[3]Indri search engine package: http://www.lemurproject.org/indri/.

[4] http://www.wordgumbo.com/ie/cel/wel/ew.htm.

[5] http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm.

[6] http://sentiwordnet.isti.cnr.it/download_1.0/.

[7] Hui Yang , Luo Si, Jamie Callan. (2006). "Knowledge Transfer and Opinion Detection in the TREC2006 Blog Track". In Proceedings of Text Retrieval Conference (TREC).