ISCAS at TREC-2004: HARD Track

Le Sun Junlin Zhang Yufang Sun

Institute of Software, Chinese Academy of Sciences
Beijing, 100080
{sunle, junlin, yufang}@iscas.cn

Abstract

Institute of Software, Chinese Academy of Sciences (ISCAS) participated in TREC-2004, submitting 18 runs. We focus on studying the problem of the combination of the user- and query-information from clarification forms and metadata. We provided two kinds of Clarification Form. Our experiment shows the CF2 is more effective than CF1. We use Google as a resource for query expansion base on metadata subject and familiarity together, and the R-prec is increased from 0.2308 (baseline) to 0.2646 (+14.6%). Our approach to exploiting the metadata Genre and Geography yield negative result when used alone, however, surprisedly, when combinate metadata Genre and metadata Geography with CF2 respectively, we get an increase (+1.2%) and (+5.4%) than use CF2 alone. Our combination of CF2 and metadata relt_text is the best results of all the TREC runs (R-prec), and in this run, the R-prec is increased from 0.3303 (CF2 alone) to 0.3766 (+14%), and from 0.2888 (metadata rel-text alone) to 0.3766 (+30.4%). From the results we can see the information from user (CF2) and the information from query (metadata relt-text) may complement each other.

1 Introduction

Institute of Software, Chinese Academy of Sciences (ISCAS) participated in TREC-2004 in all the three aspects of the HARD task. We focus on studying the problem of the combination of the user- and query-information from clarification forms and metadata. We provided two kinds of clarification form. One is a term list; the other is a title list. Metadata subject and familiarity are used together base on Google. Feature term lists are built for Metadata Genre and Geography respectively. For passage retrieval, we first cut these documents into small pieces and then do the same run as document retrieval. We totally submitted 18 results that are constructed automatically though some of them are time consuming. The following subsections describe the system design for each of the runs.

2. System Description

2.1 IR model

We focus our research on the using of the CF and metadata, so no much work is done about the IR

model. VSM is employed to calculate the similarity between query vector and document vector[1,2,3,4,5]. The term of vector is word. If T={ t_j } is a term set, then query vector v_j of topic j can be express $V_j=(v_{j1},v_{j2},\ldots v_{jn})$, in which v_{jk} denotes the weight of t_k in v_j. The vector $D_i=(d_{i1},d_{i2},\ldots,d_{in})$ denotes a document, d_{ik} denotes the weight of t_k in d_i. The similarity between v_j and d_i is calculated by following formula:

$$s_j = \sum_{k=1}^{n} d_{ik} * v_{jk} \Big/ \sqrt{\sum d_{ik}^2 + \sum v_{jk}^2}$$

The IR system we used is the Lemur toolkit (version 2.2) developed by CMU [6]. The Lemur Toolkit is designed to facilitate research in language modeling and information retrieval, where IR is broadly interpreted to include such technologies as ad hoc and distributed retrieval, cross-language IR, summarization, filtering, and classification. The toolkit supports indexing of large-scale text databases, the construction of simple language models for documents, queries, or sub-collections, and the implementation of retrieval systems based on language models as well as a variety of other retrieval models. The system is written in the C and C++ languages, and is designed as a research system to run under Unix operating systems.

2.2 Clarification Forms

We provided two kinds of clarification form. One is a list of keywords that might appear in relevant documents, the other is a list of the title and keywords of the top 10 relevant documents. The user is asked to judge the keywords or the title and keywords for a query as being "Good", "Bad", or "Unknown". The user can also input additional terms that help define or disambiguate the topic area for a query. The clarification form CHAS1 and CHAS2 are show in the following figure 1 and 2.

Figure 1: CHAS1 Clarification Form for Query Hard-409

Figure 2 CHCHAS2 Clarification Form for Query Hard-403

2.3 Exploiting Metadata

Subject & Familiarity

We use metadata subject & Familiarity together base on Google. Google is used as a resource for query expansion. First input the title of a query to Google. Then choose the related web site from the top 3. If the sites' Google directory is classified the same as the subject show at metadata of the query, then these web sites are related sites for the query. As a special case, if none of the top 3 sites is classified the exactly same as the metadata subject, we only choose top 1 as related site.

How many texts within the related site should be used as related texts is decided by metadata familiarity based on an assumption, that is, the less the user is familiar with a topic, the more he wants to know. For example, if the metadata of a query is little, then use crawler to get 2 level of web pages as the resource for query expansion; if the metadata of a query is much, then only get 1 level of web pages as the resource for query expansion. The texts of the related web pages are extracted and POS tagged. Only use these frequently appeared noun words in the related texts for query expansion. Here is the example for query HARD-401, 402 and 403.

NUMBER	Topic Title	Metadata Subject	Google Directory	Http1	the level we need	Metadata Familiarity
HARD-401	Bass Amps	TECHNOLOGY	NONE	http://www.ampeg.com/	1	**Little**
HARD-402	Identity Theft	SOCIETY	SOCIETY	http://www.consumer.gov/idtheft/	2	**Much**
HARD-403	Heaven's Gate	SOCIETY	SOCIETY	http://religiousmovements.lib.virginia.edu/nrms/hgprofile.html	2	**Much**

Genre

Two feature term lists related to different Genres, named Report Feature Word List(RFWL) and

Opinion Feature Word List(OFWL), are constructed manually.

RFWL contains some words which often appear in the news reports with high frequency such as "report" , "say", " quote ". For news report usually starts by location and time information such as "local daily the New Vision reported Tuesday ", so we also add the time information like "Tuesday" "Monday" into the RFWL to represent the features of news report. OFWL contains some words which can be regarded as the features of opinion-biased document such as "editorial " , " opinion ", " perspective "," comment".

The basic assumption is that if one document contains enough feature words listed in RFWL or OFWL, it will be regarded as the news report or opinion-editorial. So we re-rank the retrieval results by adopting the following strategy:

(a) If the requirement of the Genre metadata is "news-report " in the query, we will count the feature words in the document by looking up the RFWL. When the count is above 5 in one document, we will promote the ranking position of the document (10 position higher).

(b) If the requirement of the Genre metadata is "opinion-editorial ", we will count the feature words in the document by looking up the OFWL. When the count is above 5 in one document, we will promote the ranking position of the document in ranking list (10 position higher).

(c) If the requirement of the Genre metadata is "other ", we will punish the document that belongs to "news-report" or " opinion-editorial" by decreasing the ranking position of the document in ranking list (10 position lower).

(d) If the requirement of the Genre metadata is "any", we will keep the original ranking list without any position promotion or decreasing.

Geography:

In order to make use of the Geography metadata in queries, we construct a list that contains the main U.S. state names and some big city names. The basic assumption is that the topic of the document is relative with U.S. if it contains many U.S. state or city names. So we re-rank the retrieval results by adopting the following strategy:

(a) If the requirement of the Geography metadata is "U.S." in the query, we will count the city names and state names in the document by looking up the name list. When the count is above 5 in one document, we will promote the ranking position of the document (10 position higher).

(b) If the requirement of the Geography metadata is "non-U.S.", we will count the city names and state names in the document by looking up the name list. When the count is above 5 in one document, we will decrease the ranking position of the document in ranking list (10 position lower).

(c) If the requirement of the Geography metadata is "any", we will keep the original ranking list without any position promotion or decreasing.

Related Text: The relevant texts are used as the basis for automatic query expansion. A POS tagger is used and only high frequency noun words (except stop words) in the related text are used for query expansion.

3. Results Analysis

We submitted 1 baseline run and 18 other runs all together. We combined all the above information by different ways and wish could get more accurate ranked lists. The details of our submission experiments are show in table 1 that denotes an experiment and how the original query was constructed.

Table 1 submission experiment's detail

RUN ID	CF	MetaData	Relt Texts	Granularity	Note
ISCAS_0	no	no	no	Document	Baselin-1
Chastdn_1	no	no	no	Document/Passage	Baseline-2
Chascfw_2	CF1	no	no	Document/Passage	
Chascfd_3	CF2	no	no	Document/Passage	
Chascfwd_4	CF2	no	no	Document/Passage	Same as Chascfd_3, Only different at word weight
Chasbsubfam_5	no	Subject&Familiarity	no	Document/Passage	Based on Google
Chasbcsubfam_6	no	Subject&Familiarity	no	Document/Ptassage	CF2 is used to decide to use Google or not
Chascfsubfam_7	CF2	Subject&Familiarity	no	Document/Passage	Same as Chasbsubfam_5, Only different at word weight
Chasccsubfam_8	CF2	Subject&Familiarity	no	Document/Passage	Same as Chasbsubfam_6, Only different at word weight
Chasbaserel_9	no	no	yes	Document/Passage	Only use noun word in relt texts
Chasbasegen_10	no	Genre	no	Document/Passage	
Chasbaseger_11	no	Geography	no	Document/Passage	
Chascfrel_12	CF2	no	yes	Document/Passage	Only use noun word in relt texts
Chascfgen_13	CF2	Genre	no	Document/Passage	
Chascfger_14	CF2	Geography	no	Document/Passage	
Chasregenger_15	CF2	Genre & Geography	yes	Document/Passage	
Chasdcfd_16	CF2	no	no	Document	
Chasdcfwd_17	CF2	no	no	Document	Same as Chasdcfd_16, Only different at word weight
Chasdcfw_18	CF1	no	no	Document	

In table 2 the R-Prec and Avg Prec of our submission runs for document level are showed. ISCAS_0 is our baseline-1 run and no query expansion is used. There are two differences between ISCAS_0 and Chastdn_1. One is ISCAS_0 only use title section of a query but Chastdn_1 use both title and description section of a query. The other is ISCAS_0 is document level but Chastdn_1 is document and passage level. In the following experiments, we always use title and description section of a query as input.

Table 2 tthe R-Prec and Avg Prec of our submission run for Doc level

RUN ID	R-Prec (Hard-rel)	Avg Prec (Hard-rel)	R-Prec (Soft_rel)	Avg Prec (Soft_rel)
ISCAS_0	0.2295	0.2374	0.2716	0.2474
Chastdn_1	0.2308	0.2169	0.2772	0.2246
Chascfw_2	0.2569	0.2339	0.2745	0.2423
Chascfd_3	0.3303	0.3032	0.3225	0.2907
Chascfwd_4	0.3397	0.3161	0.3388	0.3012
Chasbsubfam_5	0.2646	0.2435	0.2808	0.2466
Chasbcsubfam_6	0.2676	0.2567	0.2959	0.2600
Chascfsubfam_7	0.3416	0.3261	0.3485	0.3175
Chasccsubfam_8	0.3423	0.3244	0.3484	0.3151
Chasbaserel_9	0.2888	0.2742	0.2822	0.2560
Chasbasegen_10	0.1913	0.1944	0.2416	0.2026
Chasbaseger_11	0.2169	0.2119	0.2480	0.2114
Chascfrel_12	**0.3766**	0.3588	**0.3717**	0.3442
Chascfgen_13	0.3355	0.3080	0.3303	0.2923
Chascfger_14	0.3483	0.3132	0.3402	0.2965
Chasregenger_15	0.3710	0.3485	0.3616	0.3348
Chasdcfd_16	0.3438	0.3254	0.3410	0.3184
Chasdcfwd_17	0.3485	0.3345	0.3509	0.3240
Chasdcfw_18	0.3049	0.2963	0.2988	0.2885
TREC median	**0.2690**	**0.2617**	**0.2906**	**0.2634**
TREC max	**0.3766**	**0.3635**	**0.3717**	**0.3554**

As we can see from table 2, the R-prec of Chasfw_2 is increased from 0.2308 (Chastdn_1, baseline-2) to 0.2569 (+11.3%) for use CF1 as query expansion, and the R-prec of Chascfd_3 is increased from 0.2308 to 0.3303 (+43.1%) for use CF2 as query expansion. So the CF2 is more effective than CF1. There is no significant improvement between Chascfd_3 and Chascfwd_4 for the only difference is word weight at query expansion.

At Chasbsubfam_5 we use Google as a resource for query expansion base on metadata subject and familiarity together, and the R-prec is increased from 0.2308 to 0.2646 (+14.6%). At Chasbcsubfam_6, we use CF2 to decide to use Google as a resource for query expansion or not, that is, if there are above 5 documents at CF2 which are denoted by user as good, then no query expansion is used. We can see from table 2 there is no significant R-prec improvement between Chassubfam_5 and Chasbcsubfam_6. At Chascfsubfam_7 and Chasccsubfam_8, we use both CF2 and metadata subject & familiarity as query expansion, and we can see the R-prec of Chascfsubfam_7 is increased from 0.3303 (Chascfd_3, CF2 alone) to 0.3416 (+3.4%).

At Chasbaserel_9, we only use metadata related text as query expansion, and get an improvement from 0.2308 to 0.2888 (+25.1%) for R-prec.

We use metadata Genre and metadata Geography respectively at Chasbasegen_10 and Chasbaseger_11, and get a decrease of R-prec from 0.2308 to 0.1913 (-17.1%) and from 0.2308 to 0.2119 (-8.2%). So our approach to exploiting the metadata Genre and Geography yield negative result when used alone.

As we can see from table 2 the chasfrel_12 is the best of all the TREC runs. At this run, we use both the CF2 and metadata relt_text as the base for query expansion. The combination of the user-information (CF2) and query-information (metadata relt_text) get significant improvement, and the R-prec is increased from 0.3303 (Chascfd_3, CF2 alone) to 0.3766 (+14%), and from 0.2888 (Chasbaserel_9, metadata rel-text alone) to 0.3766 (+30.4%). From the results we can see the information from user (CF2) and the information from query (metadata relt-text) may complement each other.

At Chascfgen_13, we combinate the user- information (CF2) and query-information (metadata Genre), and get a little increase from 0.3303 (Chascfd_3, CF2 alone) to 0.3355 (+1.2%). Similarly, at Chascfger_14, we combinate the user- information (CF2) and query-information (metadata Geography), and get an increase from 0.3303 (Chascfd_3, CF2 alone) to 0.3483 (+5.4%). Finally, we combinate the user- information (CF2) and query-information (metadata rel-text, metadata Genre, metadata Geography) and get an improvement for R-prec from 0.3483 (Chascfger_14) to 0.3710 (+6.5%).

In figure 1, the comparison of TREC Median and Maximum to ISCAS's submitted baseline run (ISCAS_0) and ISCAS's best submitted run (chascfrel_12) is showed.

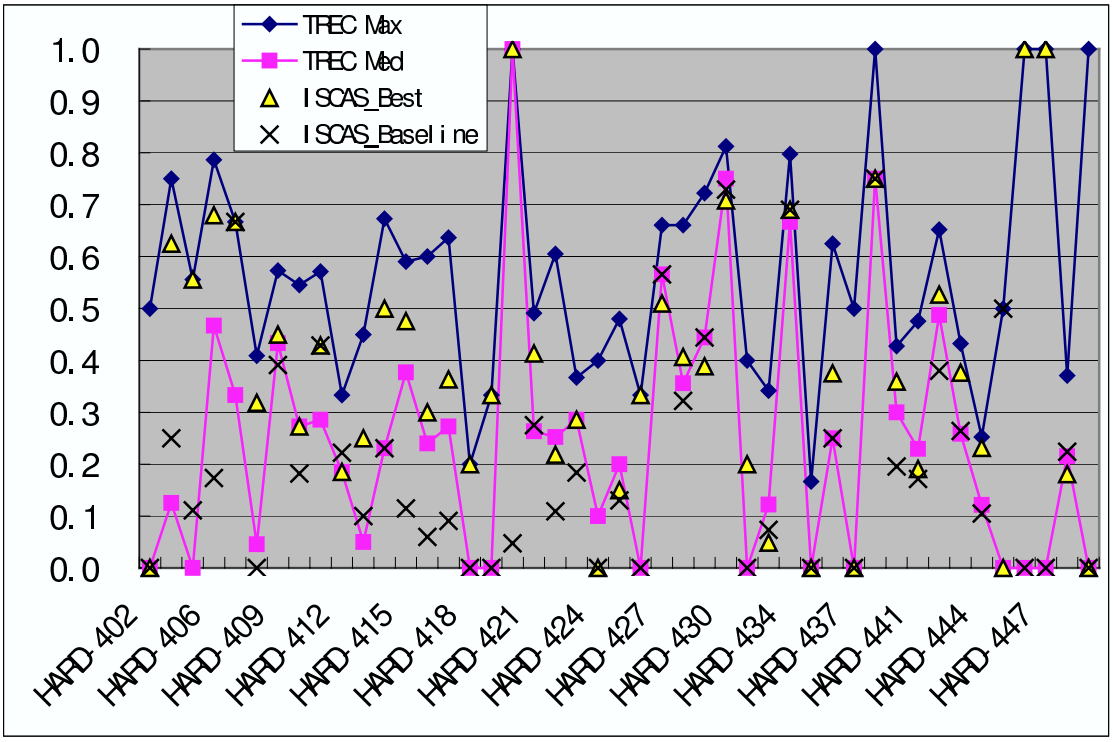

Figure 1 Comparison of TREC Median and Maximum to ISCAS's submitted baseline run and ISCAS's best submitted run (chascfrel).

4 Conclusions

We participated in all the three aspects of the HARD task and focus on studying the problem of the combination of the user- and query-information from clarification forms and metadata. We totally submitted 18 results that are constructed automatically. We provided two kinds of clarification form. CF1 is a list of keywords that might appear in relevant documents, and CF2 is a list of the title and keywords of the top 10 relevant documents. From our experiment, we can see the CF2 is more effective than CF1. We use Google as a resource for query expansion base on metadata subject and familiarity together, and the R-prec is increased from 0.2308 (baseline) to 0.2646 (+14.6%).

It's seems our approach to exploiting the metadata Genre (R-prec -17.1%) and Geography (R-prec -8.2%) yield negative result when used alone, however, surprisedly, when combinate metadata Genre and metadata Geography with CF2 respectively we get an increase (+1.2%) and (+5.4%) than CF2 alone.

Our combination of the user- information (CF2) and query-information (metadata relt_text) is the best results of all the TREC runs (R-prec), and in this run, the R-prec is increased from 0.3303 (CF2 alone) to 0.3766 (+14%), and from 0.2888 (metadata rel-text alone) to 0.3766 (+30.4%). From the results we can see the information from user (CF2) and the information from query (metadata relt-text) may complement each other.

More experiments and analysis are needed in near future.

Acknowledgements
Partly supported by the National Natural Science Foundation of China under Grant No.60203007 and the new star plan of science & technology of Beijing under Grant No.H020820790130.

References

[1] Salton,G.(1971).The SMART Retrieval System.Englewood Cliffs, N, J, Prentice-Hall, Inc.
[2] Salton,G. and Lesk,M.(1971). Computer evaluation of indexing and text precessing.Prentice-Hall,pp.143-180.
[3] G..Salton(1971).The SMART Retrieval System-Experiments in Automatic Document Processing. Englewood Cliffs,Prentice Hall,1971.
[4] G..Salton(1972).A new comparison between conventional indexing and automatic text processing(SMART). Journal of the American Society for Information Science.23(1).pp.75-84..
[5] Salton,G.(1983).Introduction to Modern Information Retrieval, McGraw-Hill
[6] http://www-2.cs.cmu.edu/~lemur/
[7] J. Allan, HARD Track Overview in TREC 2003:High Accuracy Retrieval from Documents, the proceeding of Twelfth Text Retrieval Conference (TREC 2003)
[8] L. Grunfeld, K.L. Kwok, N. Dinstl, P. Deng, TREC 2003 Robust, HARD and QA Track Experiments using PIRCS, the proceeding of Twelfth Text Retrieval Conference (TREC 2003)