NTU Approaches to Subtopic Mining and Document Ranking at NTCIR-9 Intent Task

Chieh-Jen Wang, Yung-Wei Lin, *Ming-Feng Tsai and Hsin-Hsi Chen
Department of Computer Science and Information Engineering,
National Taiwan University
*Department of Computer Science, National Chengchi University,
Taipei, Taiwan
{cjwang, ywlin, mftsai}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

ABSTRACT

Users express their information needs in terms of queries to find the relevant documents on the web. However, users' queries are usually short, so that search engines may not have enough information to determine their exact intents. How to diversify web search results to cover users' possible intents as wide as possible is an important research issue. In this paper, we will propose several subtopic mining approaches and show how to diversify the search results by the mined subtopics. For Subtopic Mining subtask, we explore various subtopic mining algorithms that mine subtopics of a query from enormous documents on the web. For Document Ranking subtask, we propose re-ranking algorithms that keep the top-ranked results to contain as many popular subtopics as possible. The re-ranking algorithms apply sub-topics mined from subtopic mining algorithms to diversify the search results. The best performance of our system achieves an I-rec@10 (Intent Recall) of 0.4683, a *D-nDCG*@10 of 0.6546 and a *D#-nDCG*@10 of 0.5615 on Chinese Subtopic Mining subtask of NTCIR-9 Intent task and an I-rec@10 of 0.6180, a *D-nDCG*@10 of 0.3314 and a *D#-nDCG*@10 of 0.4747 on Chinese Document Ranking subtask of NTCIR-9 Intent task. Besides, the best performance of our system achieves an I-rec@10 of 0.4442, a *D-nDCG*@10 of 0.4244 and a *D#-nDCG*@10 of 0.4343 on Japanese Subtopic Mining subtask of NTCIR-9 Intent task and an I-rec@10 of 0.5975, a *D-nDCG*@10 of 0.2953 and a *D#-nDCG*@10 of 0.4464 on Japanese Document Ranking subtask.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Diversified Retrieval, Subtopic Mining, Search Result Re-ranking

1. INTRODUCTION

Web search engine provides an important mechanism for users to meet their information needs, which are usually formulated by queries. A query is represented by a set of keywords and is classified into two types at TREC diversity task, one is "faceted" and the other is "ambiguous". The search intent of faceted queries is usually clear, so that the search engine can report good quality results. However, information retrieval systems often fail to capture users' search intents exactly if a submitted query is ambiguous. Because an ambiguous query usually refers to multiple categories or has more than one interpretation, retrieving a list of diversified results that cover different subtopics become a feasible solution for such ambiguous queries, especially for the case without any further information.

As mentioned as above, exploring how many subtopics of a query may be interpreted is an issue. For example, the keyword "apple" may refer to three subtopics: (1) a kind of fruit, (2) a media company, and (3) a computer company. Each subtopic may contain several sub-sub-topics, for example, "iPod", "iPhone", and "iPad" belong to a computer company subtopic. Therefore, the subtopic of a query is in terms of a hierarchal structure and contains top-down relationships. After exploring subtopics of a query, search engine can rank documents to include various subtopics at the first report page, so that users can quickly find their interesting results. However, diversity and relevance are a trade-off, because users usually concern only some specific subtopics. If search results contain many rare subtopics, the users need to spend much time to search relevant documents on the report pages.

In this paper, we aim to mine subtopics as many as possible. On the one hand, too many subtopics may provide diversified information, and thus introduce many irrelevant documents. That becomes a relevance issue. On the other hand, if only the similarity between documents and a query is considered, we may retrieve many relevant documents of the same topics. That becomes a diversity issue. Therefore, how to trade-off the diversity and the relevance issues is important. We construct an information retrieval system that can keep the quality of results and allow the top-ranked results to contain multiple important subtopics. We propose two models for Subtopic Mining subtask of NTCIR-9 Intent task, such as clustering based model and related search based model. Moreover, we propose two re-ranking models to re-rank the search results based on the subtopics mined from our subtopic mining models for Document Ranking subtask of NTCIR-9 Intent task.

The rest of this paper is organized as follows. The related work is presented and compared in Section 2. The experimental dataset used in this study is described in Section 3. The subtopic mining models and the document ranking model are introduced in Section 4. Experimental results are shown in Section 5. Lastly, Section 6 concludes the remarks.

2. RELATED WORK

Queries are usually short and even ambiguous. To realize the meanings of queries, researchers define taxonomies and classify queries into predefined categories. At the query level, Broder [3] divided query intent into navigational, informational and transactional types. Nguyen and Kan [7] characterized queries along four general facets of ambiguity, authority, temporal sensitivity and spatial sensitivity. Manshadi and Li [6] classified queries into finer categories. At the session level, Radlinski and Joachims [8] mined intent from query chains and used it for learning to rank algorithm. Boldi *et al.* [2] created graphs with query phrase nodes and used them for query recommendation.

Diversifying search results have been studied at various levels and applied to different applications. The importance of diversifying the retrieved results has been recognized in the work [14]. The main idea of the work is that the relevance of a document depends not only on itself, but also on its relations with other documents. Yue and Joachims [13] formulate such a task as the problem of predicting diverse subsets; in particular, they propose a machine learning model based on maximizing word coverage. Agrawal et al. [1] present a systematic approach to diversifying search results that aim to minimize the risk of dissatisfaction of the average user. Raei et al. [9] regard the problem of diversifying results as expectation maximization and conduct experiments on query log, in an attempt to broaden the coverage of the retrieved results via user history. Santos et al. [11] use query formulation for the purpose of diversifying search results. Among the related work, this work in [4] is the most similar to ours, in which four types of data sources, including anchor texts, query logs, search results clusters, and web sites, are employed for the diversity task. In addition to the four resources, our approach also exploits the external knowledge sources such as related search information of commercial search engines, and analyzes their effects on diversified retrieval.

3. DIVERSIFIED RETRIEVAL SYSTEM

Figure 1 shows the proposed framework for a diversified retrieval system, which contains two main subtasks such as subtopic mining and document ranking. Two models are explored in the subtopic mining to generate the subtopics for a given query. The first model mines subtopics from documents itself by document clustering and the second model applies several external resources, including the related search information provided by three commercial search engines (Google, Yahoo, and Bing). After generating subtopics of a given query, the document ranking model will re-rank the search results according to subtopics mined from subtopic mining models. In the following sections, we describe how to obtain subtopics for a given query from the two different models and how to re-rank the search results.

3.1 Subtopic Mining

This section introduces two models for subtopic mining, including clustering based model and related search based models.

3.1.1 Clustering Based Model

For document clustering, we use a clustering algorithm to discover the subtopics within the initial retrieved results for a query. Features are extracted from the retrieved results and the weight of a feature is determined by *tf-idf* as follows.

$$w_{i,d} = (0.5 + \frac{0.5\,freq_{i,d}}{\max_d freq}) \times log \frac{N}{n_i}$$

Where $freq_{i,d}$ is the frequency of feature i in document d, $\max_d$ *freq* is the maximum feature frequency in document *d, N* is the total number of documents, and n_i is the number of documents in which feature *i* appears.

The K-means clustering algorithm [5] is performed on the retrieved results and the documents of similar intent are put together in a cluster. For each term in a cluster, its *tf-idf* value is calculated. The term of the highest *tf-idf* value in a cluster is selected as a subtopic of the query.

3.1.2 Related Search Based Model

Most commercial search engines provide the related search mechanisms based on their query logs, which record users' searching and browsing behaviors. The related search mechanism provides external knowledge sources to subtopic mining. Related search query is expanded from the original query. The expanded query describes an information need more precisely based on global user search behaviors recorded in the query logs. Given a query, we collect the related search queries and each related search query is regarded as a subtopic.

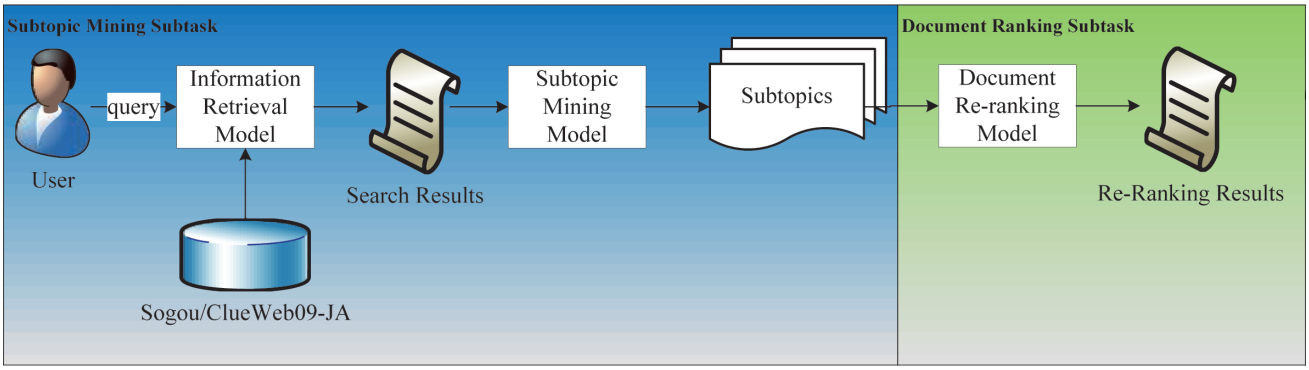

Figure 1 Framework of a Diversified Retrieval System

3.2 Document Ranking

This section introduces two diversification algorithms for re-ranking the original retrieved results, including round-robin and subtopic for diversification.

3.2.1 Round-Robin for Diversification

Round-robin is a well-known merging algorithm for merging results of various retrieval models. The main idea is to select the highest relevant documents from each subtopic, and then combine the selected documents as a final ranking list. Given a query, there are four major steps in our experiments:

- Top *n* documents are retrieved by Indri search engine.
- The K-means clustering algorithm is used to arrange the *n* documents into appropriate clusters in terms of subtopics.
- The clusters are arranged in the descending order based on their size.
- The most relevant document is selected from each cluster in a round way from larger clusters to smaller clusters.

3.2.2 Subtopic for Diversification

A document which belongs to more than one subtopic tends to have higher probability to satisfy different users, thus it is preferred. Round-robin algorithm does not allow a document in more than one subtopic. We propose a greed algorithm which integrates clues of various resources to quantify both relevance and diversity. The *Rel* function measures a document by relevance and the *Div* function estimates a document by diversity. The *Div* function gives a penalty to a document being considered if it has the same subtopics with the documents which have been selected. Given a query *q* and an original rank list R_q for query *q*, these two functions are defined as follows.

$$Rel = \frac{1}{rank(d, R_q)}$$

$$Div = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{rank(d, R_{s_{i,j}})}(1-\alpha)^{\sum_{d' \in L}\frac{1}{rank(d', R_{s_{i,j}})}}$$

Where *m* is number of subtopic mining algorithms being used, n_i is number of subtopics generated by subtopic mining algorithm *i*, $rank(d, R_q)$ returns the rank of document *d* in the retrieval result list of query *q*, $rank(d, R_{s_{i,j}})$ returns the rank of document *d* in the document list which belongs to subtopic $R_{s_{i,j}}$ (i.e., the *j*-th subtopic generated by subtopic mining algorithm *i*, and $R_{s_{i,j}}$ will be set to a very large value when *d* is not in $R_{s_{i,j}}$), *L* is a set of documents which have been selected and α is a penalty value for duplicate subtopics (0≦α≦1).

Finally, we linearly combine *Rel* and *Div* with a subtopic diversification function as follows where *ρ* is a weight of document relevance. If the parameter *ρ* equals to 1, then the ranking score will be generated according to the relevance part only.

$$\arg\max_{d \in R_q}[\rho Rel + (1-\rho)Div]$$

Algorithm 1 is used to generate a ranking list by subtopic diversification. For each iteration, the algorithm attempts to select the documents that can maximize the relevance and diversify of the final ranking list. The selected pages have to cover as many subtopics as possible. As mentioned before, the subtopics are mined by the related search model. We utilize three commercial search engines for the reference searches. After obtaining the related search query (i.e., subtopics), we query a search engine again using the related search queries. Each subtopic gets a subtopic ranking list. URLs of the original retrieved results for a given query are matched with those URLs in each subtopic ranking list. If a URL appears in the corresponding ranking list, then the document is assigned to the subtopic. Note that a document may be placed into more than one subtopic list.

Algorithm 1. A subtopic diversification algorithm

```
Input: A query $q$, a set of retrieved document $Rq$, a rank list of
       subtopics $R_s$
Output: a set of documents $L$
$L \leftarrow \varnothing$
while $|L| < 1000$
    $d\_max \leftarrow \arg\max_{d \in R_q}[\rho Rel + (1-\rho)Div]$
    $L \leftarrow L \cup \{d\_max\}$
    $Rq \leftarrow Rq - \{d\_max\}$
end while
return $L$
```

4. EXPERIMENTS

We participate in the Subtopic Mining and the Document Ranking subtasks in NTCIR-9 Intent task. We submit 3 Chinese runs to Subtopic Mining, 4 Japanese runs to Subtopic Mining, 1 Chinese run to the Document Ranking subtasks and 2 Japanese runs to the Document Ranking subtasks. The task definition and experiment results are presented in the section.

4.1 Resource Preprocessing

The two main resources adopted in this work come from Sogou[1] and ClueWeb09 collection[2]. Sogou search engine is one of major search portals in China. The query logs consist of huge user interaction collected from June 1, 2008 to June 31, 2008. Six fields are recorded in the query logs, including timestamp, userID, query content, click rank, click order, and clicked URL. In addition, Sogou provides some related resources such as clicked URL content, query type (navigational/informational), hyperlink graph data (in-link and out-link) and PageRank score of each clicked URL. The size of raw data is about 500 GB after 7z compression. In this study, Clicked URL content are segmented by Stanford Chinese Word segmenter[3] and indexed by Indri search engine, which is very popular academic information retrieval API.

Japanese document collection ClueWeb09-JA was extracted from the ClueWeb09[4] collected by the Language Technologies Institute at Carnegie Mellon University. The ClueWeb09-JA collection,

[1] http://www.sogou.com

[2] http://lemurproject.org/clueweb09/

[3] http://nlp.stanford.edu/software/segmenter.shtml

[4] http://lemurproject.org/clueweb09/

which consists of 67.3 million web pages, was crawled between January and February 2009. Similarly, all content in the dataset are segmented by Mecab segmenter[5] and indexed by Indri search engine after stop word filtering.

4.2 Task Definition

In the Subtopic Mining subtask, systems are asked to return a ranked list of subtopic strings in response to a given query. A subtopic could be a specific interpretation of an ambiguous query (e.g. "apple computer equipments" or "apple fruit" in response to "apple") or an aspect of a faceted query (e.g. "Apple iPhone App" in response to "Apple iPhone").

The Document Ranking subtask evaluated selectively the diversified search results. Systems were expected to (i) retrieve a set of documents that covers intents as many as possible; and (ii) rank documents which are highly relevant to more popular intents at higher rank of search results.

4.3 Evaluation Metric

Experiments are evaluated by the three well-known metrics: *D-nDCG* [10] which measures overall relevance across intents and Intent Recall [1] (abbreviated as I-rec) which measures diversity. For these two metrics, we list the performance at depths (i.e., number of top ranked items to be evaluated) of l = 10, 20 and 30, in order to comprehensively examine the effect of the proposed systems. *D#-nDCG* is a linear combination of I-rec and *D-nDCG*.

4.4 Experimental Results

Table 2 shows the run description for Chinese/Japanese Subtopic Mining subtask and Table 3 presents the run for Chinese/Japanese Document Ranking subtask.

Table 2 Subtopic mining subtask run

Chinese	
Run Name	Description
NTU-S-C-1	Clustering based model
NTU-S-C-2	Related search based model via Google
NTU-S-C-3	Related search based model via Bing
Japanese	
Run Name	Description
NTU-S-J-1	Clustering based Model
NTU-S-J-2	Related search based model via Google
NTU-S-J-3	Related search based model via Bing
NTU-S-J-4	Related search based model via Yahoo

Table 3 Document ranking run

Chinese	
Run Name	Description
NTU-D-C-1	Round-robin for diversification ranking
Japanese	
Run Name	Description
NTU-D-J-1	Round-robin for diversification ranking
NTU-D-J-2	Subtopic mined from Google for diversification ranking

Tables 4 shows our system performance of various subtopic mining models on the mean intent recall (I-rec), *D-nDCG* and *D#-nDCG* values for depths of 10, 20 and 30 in Chinese subtopic mining subtask. The baseline model is clustering based model (NTU-S-C-1) with K-means clustering (K=10). As shown in the table, the related search based models outperform the clustering based model. The related search based model (NTU-S-C-2) with Google achieves the best *D#-nDCG@*10 of 0.5615, outperforming the baseline about 22%. In addition, performing on the subtopics mining from Bing search engine (NTU-S-C-3) obtains 0.5558 in terms of *D#-nDCG@*10. The conclusion, i.e., related search subtopic mining models is better than clustering based model, meets our expectation because the K-means clustering algorithm depends on parameter k. Different numbers of clusters may be generated for different queries because a query may have various interpretations if it is ambiguous. The performance is decreased if more depth is considered. It may be due to the fact that the number of intents per topic less than 10 in general. Figure 2 in the overview paper of NTCIR-9 Intent Task [12] shows the details. Therefore, the subtopic mining models may propose duplicated subtopics because the number of subtopics is less than depth of evaluation metric.

Table 5 shows the Japanese subtopic mining experiment results which are revised results in the overview paper of NTCIR-9 Intent Task [12]. The tendency is similar to Table 4. Related search based model performs better than the clustering based model. The experimental results reflect again the related search based model is quite useful on Subtopic Mining subtask.

Table 6 lists the experimental results of Chinese Document Ranking subtask. As shown in the table, with the round-robin for diversification based on subtopics minded from clustering-based model with K = 10 achieves *D#-nDCG@*10 of 0.4747.

Table 7 shows the experimental results of Japanese Document Ranking subtask. Unfortunately, our submitted runs are not included in the pools and the overview paper of NTCIR-9 Intent Task [12] explains the situation. In total, there are 3,167 documents are unjudged. The submitted runs of our team are evaluated by the pooled ground truth generated by other participated teams due to lack of resources. As shown in the table, our proposed model, the document ranking model using subtopics mined from Google outperform the round-robin based model. The subtopic based document ranking model (NTU-D-J-2) achieves the best *D#-nDCG@*10 of 0.4464, better than a baseline which is round-robin based document ranking model. Two reasons may explain the phenomenon. The first reason is the coverage of the mined subtopics for a given query. The subtopics mined by clustering based model are used in the round-robin based document ranking model. The performance of the clustering based model is lower than the related search based model under Subtopic Mining subtask. That will influence the performance of the round-robin document ranking model because the clustering based model may miss some subtopics. The second reason is that the round-robin based document ranking model does not allow a document in more than one subtopic. A document belongs to more than one subtopic which tends to have higher probability to satisfy different users and it should be at higher rank. However, the round-robin based document ranking model rank documents only considering the cluster size. Documents which belong to more than one subtopic are unable to re-rank at higher rank if it belongs to a smaller cluster.

[5] http://mecab.sourceforge.net/

5. CONLUSION AND FUTURE WORK

This paper proposes systems for Subtopic Mining and Document Ranking Subtasks in NTCIR-9 Intent Task. In Subtopic Mining subtask, clustering based and related search based approaches are employed to mining subtopics for a given query. In Document Ranking Subtask, round-robin and subtopic diversification algorithms are applied to re-rank the results retrieved by the mined subtopics. A set of experiments is carried out to verify the effectiveness of the proposed system. Future directions include how to integrate more knowledge resources into the system further, such as social information, and how to extend this work to diversify Web search results with taxonomies like Open Directory Project (ODP).

6. ACKNOWLEDGMENTS

This research was partially supported by National Science Council, Taiwan under grant NSC99-2221-E-002-167-MY3. We are also grateful to Sogou search engine company for the support of SogouQ and SogouT query logs.

Table 4 Results of Chinese subtopic mining subtask

	I-rec			*D-nDCG*			*D#-nDCG*		
Run Name	@10	@20	@30	@10	@20	@30	@10	@20	@30
NTU-S-C-1	0.4335	0.4335	0.4335	0.4836	0.3140	0.2432	0.4586	0.3738	0.3384
NTU-S-C-2	0.4683	0.4683	0.4683	0.6546	0.4242	0.3278	0.5615	0.4463	0.3980
NTU-S-C-3	0.4807	0.4807	0.4807	0.6308	0.4090	0.3163	0.5558	0.4449	0.3985

Table 5 Results of Japanese subtopic mining subtask (Revised)

	I-rec			*D-nDCG*			*D#-nDCG*		
Run Name	@10	@20	@30	@10	@20	@30	@10	@20	@30
NTU-S-J-1	0.3021	0.3021	0.3021	0.2409	0.1741	0.1522	0.2715	0.2381	0.2272
NTU-S-J-2	0.4442	0.4442	0.4442	0.4244	0.3043	0.2647	0.4343	0.3742	0.3544
NTU-S-J-3	0.4205	0.4205	0.4205	0.3913	0.2831	0.2469	0.4059	0.3518	0.3337
NTU-S-J-4	0.3935	0.3935	0.3935	0.4060	0.2904	0.2509	0.3998	0.3420	0.3222

Table 6 Results of Chinese document ranking subtask

	I-rec			*D-nDCG*			*D#-nDCG*		
Run Name	@10	@20	@30	@10	@20	@30	@10	@20	@30
NTU-D-C-1	0.6180	0.6952	0.7169	0.3314	0.3706	0.3473	0.4747	0.5329	0.5321

Table 7 Results of Japanese document ranking subtask

	I-rec			*D-nDCG*			*D#-nDCG*		
Run Name	@10	@20	@30	@10	@20	@30	@10	@20	@30
NTU-D-J-1	0.5819	0.6936	0.7290	0.2426	0.2530	0.2582	0.4122	0.4733	0.4936
NTU-D-J-2	0.5975	0.6817	0.7255	0.2953	0.2825	0.2743	0.4464	0.4821	0.4999

7. REFERENCES

[1] Agrawal, R., Gollapudi, S., Halverson, A. and Ieong, S. 2009. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, 5-14.

[2] Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A. and Vigna, S. 2008. The query-flow graph: model and applications. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, 609-618.

[3] Broder, A. 2002. A taxonomy of web search. *SIGIR Forum*. 36, 2 (2002), 3-10.

[4] Dou, Z., Hu, S., Chen, K., Song, R. and Wen, J. 2011. Multi-dimensional search result diversification. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining - WSDM '11*, 475-484.

[5] Macqueen, J.B. 1967. Some Methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Math, Statistics, and Probability*, 281-297.

[6] Manshadi, M. and Li, X. 2009. Semantic tagging of web search queries. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, 861-869.

[7] Nguyen, V. and Kan, M. 2007. Functional Faceted Web Query Analysis. *Query Log Analysis: Social And Technological Challenges. A Workshop at the 16th International World Wide Web Conference*.

[8] Radlinski, F. and Joachims, T. 2005. Query chains: learning to rank from implicit feedback. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 239 – 248.

[9] Rafiei, D., Bharat, K. and Shukla, A. 2010. Diversifying web search results. In *Proceedings of the 19th International Conference on World Wide Web* , 781-790.

[10] Sakai, T. and Song, R. 2011. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1043-1052.

[11] Santos, R.L., Macdonald, C. and Ounis, I. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international Conference on World Wide Web*, 881-890.

[12] Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M., Wang, Q. and Qrii, N. 2011. Overview of the NTCIR-9 INTENT Task. In *Proceedings of the 9th NTCIR Workshop Meeting*.

[13] Yue, Y. and Joachims, T. 2008. Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th International Conference on Machine Learning*, 1224-1231.

[14] Zhai, C.X., Cohen, W.W. and Lafferty, J. 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 10-17.