HIT^2 Joint NLP Lab at the NTCIR-9 Intent Task

Dongqing Xiao[1] Haoliang Qi[2] Jingbin Gao[1]
Zhongyuan Han[1,2] Muyun Yang[1] Sheng Li[1]
[1]Harbin Institute of Technology, Harbin, China [2]Heilongjiang Institute of Technology, Harbin, China
dqxiao@mtlab.hit.edu.cn haoliang.qi@gmail.com jbgao@mtlab.hit.edu.cn

ABSTRACT
The report hereby is to represent the principle, the searching process and experiment results. We report our systems and experiments in the intent task of NTCIR 9. The research aims at evaluating the effectiveness of the proposed methods on query intent mining and results diversification in terms of web search. In the subtopic mining subtask, we combine the extracted candidates from search logs and Wikipedia. An improvement could be seen after incorporating query intents from different resources. In the document ranking subtask, greedy algorithms are taken to select documents with the high diversified score and return a re-ranked list of diversified documents based on query subtopics. The experiment results show that the method, that is combining subtopic results directly, outperforms MMR.

Keywords: subtopic mining, query logs, Wikipedia, MMR, diversification.

1. INTRODUCTION
Users are used to using Web search engines to find information. They tend to express their intentions by submitting a short query which can usually represent a variety of meanings. There are two reasons leading to this result: one is that some terms in the query are ambiguous (e.g., 'beatles' refers to a band or a kind of animal), the other is that the query can cover multiple aspects (e.g., beatles band). In the first case, the query is open to different interpretations (e.g., band, animal), and in the second case, the user might be interested in different aspects underlying the query (e.g., history, albums and songs download) [1].

Given a query, an information retrieval system should respond with a ranked list that respects both the breadth of available information and any ambiguity inherent in the query [2, 3]. This kind of approach diversifying the results has two advantages: the list can not only cover the user's current intents with the greatest probability, but also provide a choice to find a given document relevant to their information need as soon as the need already satisfied by other documents have been observed. In order to diversify the search results, a search engine should consider the relevance of each individual document. At the same time, whether different aspects under the same query can be covered should also be taken into account. By doing so, the retrieved documents should provide the maximum coverage and minimum redundancy with respect to the possible aspects underlying a query [4]. There have been many studies on diversification in the field of information retrieval [5, 6]. Most previous approaches to this problem are based on a greedy approximation algorithm, inspired by the notion of maximal marginal relevance [7]. In common, they seek to promote diversity by directly comparing the documents retrieved from a given query with one another, so as to iteratively select those results most relevant to the query as well as most dissimilar to the documents. Therefore, these approaches implicitly assume that similar documents will cover similar interpretations or aspects underlying the query, and hence should be demoted, so as to achieve a diversified ranking. However, this method is not able to effectively identify the user's intents. When a document covered a number of aspects of a query, it should be ranked on top of diversified search results, rather than simply considering its similarity with other results. As the broad topic underlying an ambiguous or underspecified query can be usually decomposed into its constituent subtopics. We can explicitly account for different aspects of the query, and make use of these subtopics to produce a diverse ranking of results.

Query intent classification contains identifying the underlying goal of the user when submitting one particular query. Since a user's query may implicitly express more than one intent, it would be very helpful if a general search engine could detect all the query intents, distribute the query to appropriate vertical search engines as well as effectively organize the results from the different vertical search engines to satisfy a user's information need. As it is crucial for providing better diversified search results, understanding a user's query intent thus can increase user's sense of satisfaction.

A variety of related query intent identification problems have been investigated in the past. One kind of strategies focuses on the associative document retrieval method in which the query is a document including multiple subtopics. Another one is to reexamine the query intent by analyzing the query logs. In addition, there is also an approach can discover large quantities of intent concepts by leveraging the third-party resources. In our method, we use the last two kinds of data mentioned above.

This paper is organized as follows. In Section 2, we describe our proposed methods for query intent mining. Section 3 provides a detailed description of document reranking task. We report experiment results in section 4, and conclude in Section 5.

2. Subtopic Mining Task
2.1 Introduction
The challenges in the subtopic mining task tie in how to extract the subtopics of the original query and to detect their importance automatically. Query subtopics are the representative information needs associated with the query [8]. Query logs realistically reflect diverse information needs, and meanwhile Wikipedia contains different interpretations of ambiguous topics and parts of user interests. In order to cope with the challenges, we mainly explore how to mine query intents from query logs and external resource and how to incorporate them into a single list.

2.2 Inferring query intent from query logs
Different user may be interested in different aspects of the general query. Users sometimes type in different queries which contain

general query terms. After entering a query and being presented with results, web search engine user may follow with query reformulations which provide alternative phrasing of user's intents, e.g. when the user was not satisfied with the presented results (they want to see more focused or more results).

According to above-mentioned analysis, given an input query *q*, the expanded query includes original query terms and reformulated query terms, which coexist in one session with the original query to find relevant query intents. We use SogouQ query log to help us to find out potential user intent or sub-queries.

To begin with, we make a census on the entire query log data. We treat each Chinese word as a vocabulary. For given original query q_0, we calculate the frequency of each bigram proceeds or after q_0, and preserve the top 32 frequent bigrams respectively.

Then we build two graphs respectively just as shown in Figure 1.

After constructing graphs for original query q_0, we apply Depth First Search algorithm to them to find out possible sub-queries just as discussed in Algorithm1.

The algorithm outputs the longest query intents starting from nodes without in-edges and ending in nodes without out-edges. According to its frequencies, we assign normalized score to each expanded queries. The algorithm has many limitations. It fails to deal with the loops in above-mentioned graphs, and the output query intents sometimes are so long that referring to a trivial user needs. Due to time limit, we don't improve this algorithm. Fortunately, we minimize its effect by utilizing other resources.

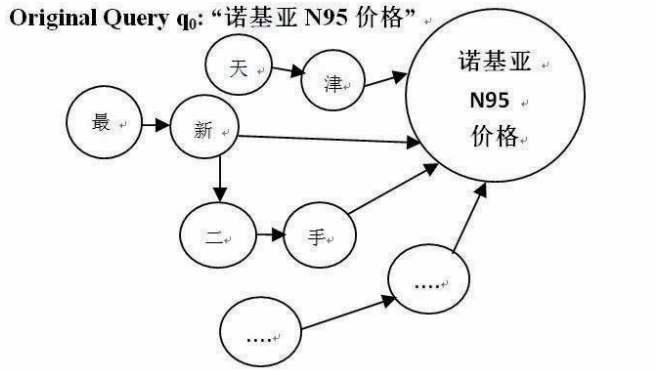

Figure 1: Example of graphs of query intents

Algorithm 1: DFS for getting query intents

```
Given G=<V,E, Frequency, in-degree, out-degree>
S=[]
For v in V and out-degree(v)=0:
  S+=v;
End for
QL=[]
For v in S:
  Qv=DFS-visit(G, v );
  QL+=Qv
End for
Calculate Cosine similarity; roughly eliminate the redundant ones
Preserve the top k=50 most frequent expanded queries
```

DFS-visit(G, v)

```
If out-degree(v)==0:
   Return v
Else:
  For v' in AdjVex(v):
    QV'=DFS-VISIT(G, v')
    For q' in QV':
      q=v + q; frequency[q]=frequency[v]*frequency[q']
      QV+=q
    End for
  End for
  Return QV
```

In this subtask, due to limited time in the official run, we don't use click logs or document dump to filter expanded queries and extract more potential query intents from them. Without effective filters, there may be some irrelevant query intents in our submitted runs. Without efficient way to diversify extracted intents, it is hard to achieve ideal lists of diverse intents solely via string similarity.

2.3 Inferring query intent from Wikipedia

In order to handle sparseness problem, we resort to external resource Wikipedia. Wikipedia is a free on-line encyclopedia edited collaboratively by large numbers of volunteers. It covers many topics, so it might be reasonably assumed to reflect diverse interpretation of each topic and parts of different user interest. As an open source project, Wikipedia is accessible via downloading from http://dumps.wikimedia.org-/zhwiki/, in form of database dumps released periodically. We download Chinese Wikipedia dump of Sep 2009.

The method we adopt for processing Wikipedia is as follows. Firstly, we retrieve Wikipedia pages by a retrieval system built based on Indri [9], then extract expanded queries from the first 20 returned pages for each topic. Intuitively, intent of initial query is a group of semantically related terms that represent a piece of relevant information of a query. Therefore, we extract terms with high frequency, titles, and named entity. To rank extracted terms, we take factors shown in Table 1into account.

Table 1: Factors indicating the relevance of terms in Wikipedia articles

Factor	Content	Description
1	Title	Unique identifiers for topic and sections
2	Overview	Lead section, summary of the topic
3	Content	Grouped by sections
4	Entity	Unique identifier for entity in Wikipedia
5	Position	Position in the whole article
6	P-position	The distance to the nearest paragraph's title

To quickly combine these factors together, we use simple linear combination of the scores obtained by scoring functions F_t as the following form:

$$P(w|D) = \frac{\sum_{t=1}^{6} w_t * F_t(w, D)}{|D|}$$

In our experiment, we set W_t and F_t just as follows.

Factor	W_t	F_t
1	1	$F_1 = \sum_{i=1}^{TF} \gamma^{\text{the level of title } (w_i, D)}$ $\gamma = 0.6$
2	0.15	$F_2 = TF_{\text{lead section}}(w, D)$
3	0.05	$F_3 = TF_{\text{content}}(w, D)$
4	0.1	$F_4 = TF_{\text{entitiy}}(w, D)$
5	1	$F_5 = e^{-\min(\text{postion}(w,D))}$
6	1	$F_6 = e^{-2*\min(P-\text{postion}(w,D))}$

2.4 Combination of two methods

As for topics covered by both resources, it is necessary to combine query intents extracted from query logs with Wikipedia to achieve better performance.

For this purpose, first, we delete the duplicate queries and filter query intents according to their lengths. In our view, a query containing more than three semantic units except original one concerns too trivial aspect, so the shorter ones are better.

Then, we normalize the probability of candidate intents mined from two resources into [0, 1] respectively.

We find that normalized probability distributions of query intents from different source are different, just as shown in figure 2. The probability of query intent mined from search logs declines quickly, while the one from Wikipedia varies gently. In addition, the number of subtopics covered by candidate terms mined from query logs ranges from 5 to 6, while, the number of subtopics covered by candidate terms mined from Wikipedia ranges from 2 to 3. Motivated by our observations, we use the following two methods to convert the score of candidate term.

Method 1:

We use simple linear combination of the normalized score to obtain final score for each term as the following form:

$$\mathrm{Score}(\mathrm{w}) = \lambda_1 \mathrm{score}_{\text{search logs}}(\mathrm{w}) + \lambda_2 \mathrm{score}_{\text{wikipedia}}(\mathrm{w})$$

$$\lambda_1 = \frac{5}{7};\ \lambda_2 = \frac{2}{7}$$

Method 2:

We transform the probability distribution of query intents mined from Wikipedia in the following way.

$$\mathrm{score}'_{\text{wikipedia}}(\mathrm{w}) = \frac{\mathrm{score}_{\text{wikipedia}}(\mathrm{w})}{2^{\mathrm{rank}(\mathrm{w})-1}}$$

$$\mathrm{Score}(\mathrm{w}) = \lambda_1 \mathrm{score}_{\text{search logs}}(\mathrm{w}) + \lambda_2\, \mathrm{score}'_{\text{wikipedia}}(\mathrm{w})$$

$$\lambda_1 = \frac{5}{7};\ \lambda_2 = \frac{2}{7}$$

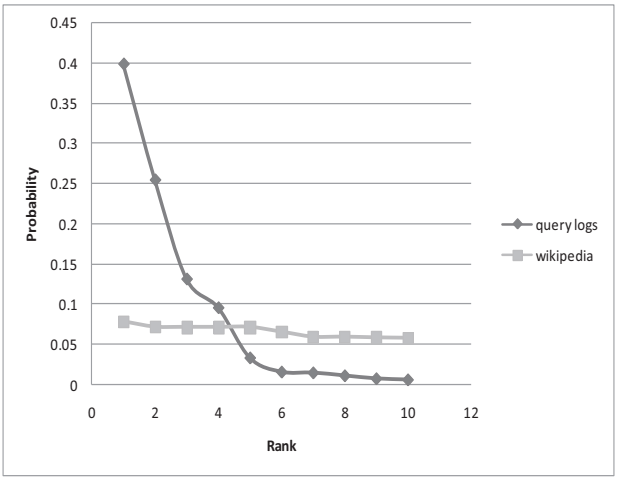

Figure 2: Example of Normalized probability distributions of query intents

To check its effectiveness, we manually label extracted query intents of 10 chosen queries according to commercial search engine query recommendation and common sense. We adopt precision and recall to evaluate the effectiveness of our method.

Table 2: Performance achieved by every method

Method	Precision@10	Intents-recall@10
Mining from Wikipedia	0.474	0.4032
Mining from search log	0.384	0.3778
Linear combination	**0.508**	**0.4486**

As can be seen in Table 2, the combination keeps passable precision while improve intent recall significantly.

3. Document Ranking Task

We use two kinds of strategies to rank the results of the queries which cover more than one subtopic. They are combination of subtopic results for the original query and diversity of the results which are generated with the previous strategy.

3.1 Introduction of MMR

The Maximal Marginal Relevance (MMR) criterion strives to reduce redundancy while maintaining query relevance in re-ranking retrieved documents. It is introduced by Jaime Carbonelland Jade Goldstein in SIGIR'98 to meet the need for "relevant novelty". It measured relevance and novelty independently and provided a linear combination as the metric. The linear combination can be called as "marginal relevance". For example, a document has high marginal relevance if it is not only relevant to the query, but also contains minimal similarity to previously selected documents. It can be seen in the following equation.

$$MMR \overset{def}{=} Arg \max_{D_i \in R \setminus S} [\lambda(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j))]$$

Where C is a document collection; Q is a query; $R = IR(C, Q, \theta)$, i.e., the ranked list of documents retrieved by an IR system, given C and Q and a relevance threshold θ, below which IR system will not retrieve documents (θ can be degree of match or number of documents); S is the subset of documents in R already selected; R\S is the set difference, i.e., the set of as yet unselected documents in R; Sim1 is the similarity metric used in document retrieval and relevance ranking between documents and a query; and Sim2 can be as same as Sim1 or be a different metric.

Given the above definition, MMR computes incrementally the standard relevance-ranked list when the parameter $\lambda = 1$, and computes a maximal diversity ranking among the documents in R when $\lambda = 0$. For intermediate values of λ in the interval [0,l], a linear combination of both criteria is optimized. Users wishing to sample the information space around the query should set λ at a smaller value, and those wishing focusing in on multiple potentially overlapping or reinforcing relevant documents should set λ to a value closer to 1.

3.2 Combination of Subtopic Results for Diversity

We retrieved all the subtopics which is excavated from one topic with our IR system, and returned 1000 documents for each subtopic. Then we combined the results of different subtopic as one document set for the original query. As the relevance between the document and the subtopic can be expressed by a score, the greater the value is the more relevant subtopic and document to each other. If they have no relevance at all, the value of the score will be considered as zero. So we calculated the relevance between the document and the original query by adding up the scores. The document with a greater value of score will be ranked higher than the one with a smaller value of score. Then we picked up the top 2000 results for each original query to introduce the

MMR method for twice diversification and retrieved the top 1000 results as the reordered results of current strategy which is marked as "HIT2jointNLPLab -D-C-2" in the table .

3.3 Integrate MMR for twice diversification

The method here we used was based on the MMR which is a classic approach for diversity.

3.3.1 Determine the parameter

We used the subtopics of 10 samples which are provided by the organizer and another 15 queries excavated by ourselves as our data set. After combining the results for each query which is mentioned in section 3.2, we tried the parameter λ from 0.1 to 0.9, and reserved the top 1000 results for each query. Then we used tool which was also provided by the organizer to evaluate the quality of the re-ranked results. As a result, we determined value of the parameter λ as 0.35.

3.3.2 Similarity measure between documents

We use the well-known Kullback-Leibler divergence (referring to KL-divergence generally) to measure the similarity of different documents which is indicated by Sim2in the MMR equation. We considered a document as a probability distribution, so the similarity of the documents can be seen as the difference between different distributions. Given the true probability distribution *Di* and another distribution *Dj* that is an approximation to *Di*, the KL divergence is defined as:

$$KL(D_i \| D_j) = \sum_x D_i(x) \log \frac{D_i(x)}{D_j(x)} .$$

Since KL-divergence is always positive and larger for the distributions that are further apart, we use the negative KL-divergence as the basis for calculating. In addition, KL-divergence is not symmetric, and it matters which distribution we pick up as the true distribution. If we assume the true distribution to be *P*, then the negative KL-divergence can be expressed as

$$\sum_{\omega \in V} P(\omega \mid D_i) \log P(\omega \mid D_j) - \sum_{\omega \in V} P(\omega \mid D_i) \log P(\omega \mid D_i) ,$$

where the summation is over all words ϖ in the vocabulary V. The second term on the right-hand side of this equation does not depend on the document *Dj*. The more close to zero the value is, the more similar the two documents are. Given a simple maximum likelihood estimate for $P(\varpi|Dj)$, based on the frequency in the document text ($f_{\varpi, Di}$) and the number of words in the document *Di* ($|Di|$), the left-hand side of this equation will be $\sum_{\omega \in V} \frac{f_{\omega, D_i}}{|D_i|} \log P(\omega \mid D_j)$.

However, the major problem with this estimate is that if any of the document Di words are missing from the document *Dj,* the $logP(\varpi| Dj)$ will lose meaning. So we have to do some smoothing which is a technique for avoiding this estimated problem and overcoming data sparseness. If $P(\varpi|C)$is the probability of word ϖ in the document collection C, then the estimate we use for an unseen word in the document *Dj* is $\alpha_D P(\varpi|C)$, where α_D is a coefficient control over the probability assigned to unseen words.

We set α_D as 0.2 here. In order to ensure the probabilities sum to one, the probability estimate of a word that is seen in a document is $(1 - \alpha_D) P(\varpi|D) + \alpha_D P(\varpi|C)$. Fortunately, we can get the $P(\varpi|C)$from our index by using the API provided by Indri in a constant time.

3.3.3 Rerank the documents

After the work we have introduced in the previous section, we can use MMR method to rerank the 2000 documents for each original query. Then we picked up the top 1000 documents as our results for each original query which was marked as "HIT2jointNLPLab -D-C-1" in the table.

4. Experiment Results

4.1 Evaluation metric

For both tasks, NTCIR used D#-nDCG as primary evaluation metrics. D#-nDCG is a linear combination of intent recall (I-recall, which measures diversity) and D-nDCG (which measures overall relevance across intents) [10]. The higher I-rec means higher intent recall, the higher D-nDCG means better overall ranking. D#-NDCG explicitly encourages high intent recall and global ideal ranking in a search output.

4.2 Experiment results on subtopic mining

Table 3: Official results of our submitted runs in subtopic mining task

run name	I-rec@10	DnDCG@10	D#-nDCG@10
HIT2jointNLPLab-S-C-1	0.4240	0.5946	0.5093
HIT2jointNLPLab-S-C-2	0.4596	0.6407	0.5501

We submit two runs in subtopic mining subtask of intent task. All of them are combination of query logs and Wikipedia.

To evaluate effectiveness of proposed methods for query intent mining and find out what effects effectiveness, we compare our performance with median one among submitted runs per topic.

Table 4：The number of queries in different categories when computing our runs with media runs of subtopic mining task based on D#-nDCG@10

run name	>median	=median	< median	Total
HIT2jointNLPLab -S-C-1	65	4	31	100
HIT2jointNLPLab -S-C-2	74	4	22	100

Table 5：The number of queries in different categories when computing our runs with media runs of subtopic mining task based on I-recall@10

run name	>median	=median	< median	Total
HIT2jointNLPLab -S-C-1	53	5	42	100
HIT2jointNLPLab -S-C-2	62	1	**37**	100

The results are based on queries that NTCIR releases the statistical information of all the submitted to subtopic mining task in Chinese collection. According to the statistics, other submitted runs slightly outperform our runs in I-recall.

Table 6: The number of queries worse than media runs in different categories considering its query intent origin.

Query	In query logs only	In Wiki only	Overlap	Omission	total
Num	10	1	**16**	**10**	**37**

Table 7: Topic coverage

Query	In query logs only	In Wiki only	Overlap	Omission	total
Num	32	6	**52**	**10**	**100**

Based on detailed per-topic analysis, we find two factors which affect our method's performance significantly.

The first one is topic coverage. As can be seen in Table 7, there are 10 topics , taken up over 27%（10/37）absent from query logs and Wikipedia. Compared with other external resources e.g. Baidu encyclopedia, Chinese Wikipedia covers fewer entities. Hence, we can improve intent recall by using other external resources which contain richer entities in Chinese.

The second one involves with the combination method. As it can be seen in Table 8, 16 topics out of 52 topics are found to be worse than the median in Intent recall. Despite that topics are covered by query logs and Wikipedia, the integrated result is still unsatisfactory. According to our observation, the frequency of intents extracted from query logs is not significantly related with its probability. Thus, during the combination, it is easy to leave out some relevant query intents with low frequency in query logs while relative high probability in official results. Our method of combining query intents from different sources is quite native. Additionally, the problem, how to detect the importance of subtopics, still needs further study.

4.3 Diversity of Results

The below tables show the mean intent recall, D-nDCG and D_#nDCG values for top l = 10, 20, 30 documents with our strategies. And our best result is ranked at 13/24, 14/24 and 15/24, respectively for different number of documents.

run name	I-rec@10	D-nDCG@10	D#-nDCG@10
HIT2jointNLPLab -D-C-2	0.5794	0.3704	0.4749
HIT2jointNLPLab -D-C-1	0.4716	0.3573	0.4144

run name	I-rec@20	D-nDCG@20	D#-nDCG@20
HIT2jointNLPLab -D-C-2	0.6815	0.3928	0.5371
HIT2jointNLPLab -D-C-1	0.5499	0.3819	0.4659

run name	I-rec@30	D-nDCG@30	D#-nDCG@30
HIT2jointNLPLab -D-C-2	0.7057	0.3656	0.5357
HIT2jointNLPLab -D-C-1	0.5752	0.3499	0.4625

From the table we can see that the result combined with subtopic results directly is better. It is reasonable if a result is related to more subtopics; it should be ranked higher in the result list. And because of the limit of the number of document collection retrieved, the role of the diversity results is hardly marked. It may have high risks of lowering the ranking of results which can cover more important subtopics. But it is necessary to perform diversity in the commercial search engine to prevent duplication.

The subtopics which are mined from user logs will make more contribution to our results than those which are extracted from Wikipedia. Because the log files implicitly contain some interaction information between users and the search engine, such information may be more useful since they can well represent users' intentions. And satisfaction with the current results means the result list can cover the user's intents better.

Compared with the approaches leading to better results, our methods missed the information of hyperlink, sites and anchor. These kinds of information can reflect some relationships among different web pages. They can help us filter some similar pages. Meanwhile, we did not consider the likelihood of each subtopic when we ranked the results. It can help to find the pages which only cover little rare subtopics.

Since the subtopic collection used as standards for evaluation covers all the subtopics submitted by different teams, -our re-ranked results will have a bad performance for the subtopics which are not covered in out method. We also did not consider how to adjust the weight to the document which covered multiple subtopics when reordering the results, but only added up all the scores together simply. However, as some subtopics are similar to each other, it is reasonable to reduce the weight of the documents that covered the similar subtopics. In our approach, we determined the similarity between documents by only considering the contents of the documents. In fact, lots of information such as URL and hyperlink related to a document should also be taken into account when calculating the similarity between different documents. In terms of the text in the document, there were many noises existing and we did not handle them well. For example, there are a lot of web advertisements in the web pages which will influence our similarity calculation. In addition, the issue that how we can make use of the document titles, text contents and sources of pages when calculating the similarity also requires more efforts.

5. Conclusion

In the subtopic mining task, different strategies for mining subtopics are applied to different resources. An improvement could be seen after combining query intents from different resources. While, more investigations are needed to explore potential of other resources in recalling diverse query intents, filtering irrelevant ones and detecting the probability of query intents. In the document ranking subtask, two kinds of methods are applied. The experiment results show that the method, that is combining subtopic results directly, outperforms MMR.

6. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. 60736014 and the National High Technology Development 863 Program of China under Grant No. 2006AA010108.

7. REFERENCES

[1] C. L. A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In Proc. of ICTIR, pages 188–199, 2009.

[2] W. S. Cooper. The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. Technical report, Univ. of California, 1971.

[3] S. E. Robertson. The probability ranking principle in IR. Journal of Documentation, 33(4):294–304, 1977.

[4] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. B¨uttcher, and I. MacKinnon. Novelty and diversity in

information retrieval evaluation. In Proc. of SIGIR, pages 659–666, 2008.

[5] B. Boyce. Beyond topicality: A two stage view of relevance and the retrieval process. Info. Processing and Management, 18(3):105-109, 1982.

[6] W. Goffman. On relevance as a measure. IP&M, 2(3):201–203, 1964.

[7] Zhang, Z. (2006). Mining Search Engine Query Logs for Query Recommendation BASED SIMILARITY FOR CONSECU-, 2-3

[8] Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 Web Track. Language, 1-9.

[9] http://www.lemurproject.org/indri/

[10] T. Sakai and R. Song. Evaluating Diversified Search Results Using Per-Intent Graded Relevance. In Proceedings of ACMSIGIR 2011, pages 1043–1052, 2011

[11] M. A. Hearst. Search User Interfaces. Cambridge University Press, 2009.

[12] K. Sp¨arck-Jones, S. E. Robertson, and M. Sanderson. Ambiguous requests: implications for retrieval tests, systems and theories. SIGIR Forum, 41(2):8–17, 2007.

[13] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In Proc. of WSDM, pages 5–14, 2009.

[14] Broder, A., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., & Zhang, T. (2007). Robust Classification of Rare Queries Using Web Knowledge Search.