Understanding the Query: THCIB and THUIS at NTCIR-10 Intent Task

Junjun Wang, Guoyu Tang, Yunqing Xia, Qiang Zhou, Fang Zheng
Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China
{jjwang, gytang, yqxia, zq-ld, fzheng}@tsinghua.edu.cn

Qinan Hu, Sen Na, Yaohai Huang
Canon Information Technology (Beijing) Co. Ltd., Beijing 100081, China
{huqinan, nasen, huangyaohai}@canon-ib.com.cn

ABSTRACT

Understanding intent underlying search query recently attracted enormous research interests. Two challenging issues are worth noting: First, words within query are usually ambiguous while query in most cases is too short to disambiguate. Second, ambiguity in some cases cannot be resolved according merely to the limited query context. It is thus demanded that the ambiguity be resolved/analyzed within context other than the query itself. This paper presents the intent mining system developed by THCIB and THUIS, which is capable of understanding English and Chinese query respectively, with four types of context: query, knowledge base, search results and user behavior statistics. The major contributions are summarized as follows: (1) Extracted from the query, concepts are used to extend the query; (2) Concepts are used to extract explicit subtopic candidates within Wikipedia. (3) LDA is applied to discover explicit subtopic candidates within search results. (4) Sense based subtopic clustering and entity analysis are conducted to cluster the subtopic candidates so as to discover the exclusive intents. (5) Intents are ranked with a unified intent ranking model. Experimental results indicate that our intent mining method is effective.

Team Name

THU+CIB/THCIB; THU/THUIS

Subtasks

Subtopic Mining (English, Chinese)

Keywords

Intent, concept, subtopic mining, subtopic ranking, word sense induction

1. INTRODUCTION

Web search engine meet information need by simply returning a ranked list of search results according to a user-specified query. There is always a certain purpose (i.e., intent) before a query is issued to a search engine, which is usually classified into three types: transactional (where the user is interested in some Web-mediated activity), navigational (where the user has a particular URL to find) and informational (in which the user has an information need to satisfy) [23]. However, the dominating informational searches are usually expressed by vague, broad or ambiguous queries. Two challenging issues are worth noting: First, words within query are usually ambiguous while query in most cases is too short to disambiguate. Second, ambiguity in some cases cannot be resolved according merely to the limited query context. It is thus demanded that the ambiguity be resolved/analyzed within context other than the query itself.

NTCIR-10 intent task brings together research efforts in addressing the above demand by defining two subtasks: subtopic mining and document ranking [27]. In the subtopic mining subtask, systems are required to return a ranked list of subtopic strings in response to a given topic query. The top N subtopic strings should be both relevant and diversified as much as possible. In the document ranking subtask, systems should return selectively diversified Web search results. The returned documents should cover intents as many as possible, and are ranked based on relevance and diversity.

The THUIS team comprised of researchers from Intelligent Search group at Tsinghua University participated in the Intent task of NTCIR-10 in Chinese. THCIB team, a joint team between THUIS and Canon Information Technology (Beijing) Co. Ltd., participated in the Intent task of NTCIR-10 in English. In the intent mining system, we concentrate on subtopic mining. The research efforts are intensively made on concept-based text analysis, which seeks to achieve understanding of both query and query-related text content. We summarize contributions of this work as follows. First, we extracted concepts from query with Wikipedia. The concepts are in turn used to extract explicit subtopic candidates within Wikipedia. Second, we applied LDA to discover explicit subtopic candidates within search results. Third, a sense-based clustering algorithm is designed to discover the exclusive intents from the subtopic candidates. Fourth, we ranked subtopics with a unified model considering both relevance and diversity.

The proposed intent mining method is established upon concepts. This makes our method theoretically advantageous over the word-based method. Fortunately, experimental results justify our claim. Two further conclusions are also interesting. First, clustering algorithm is helpful to find exclusive intents hidden in the subtopic candidates, which are crucial to subtopic ranking. Second, the proposed unified ranking model, though not outperforms the state-of-the-art relevance based ranking model, is potential to improve.

The rest of this paper is organized as follows. In Section 2, we summarize related work on intent mining and search

results clustering. In Section 3, the intent mining system is briefly described. Algorithm for subtopic mining and ranking are presented in Section 4 and Section 5, respectively. We present evaluation and discussion in Section 6, and conclude this paper in Section 7.

2. RELATED WORK

Intent mining task has been organized twice by NTCIR evaluation meeting [27, 25]. Many methods are mentioned in NTCIR-9 intent mining task reports. Meanwhile, our method is highly related to research work on search results clustering. We summarize related work within the two categories as follows.

2.1 Intent Mining

In the NTCIR-9 intent mining task, 15 systems contributed 42 runs for the subtask of Chinese subtopic mining [25]. For the Japanese subtopic mining task, 3 systems contributed 14 runs for the Japanese task [25]. We summarize the interesting work as follows.

Firstly, systems using multiple resources tend to be advantageous. For example, THUIR system uses Google, Bing, Baidu, Sogou, Youdao, Soso, Wikipedia and query log [28]. ICTIR system uses Baidu, Sogou, SoSo, Wikipedia, Hudong Encyclopedia and query log [30]. HITCSIR system uses Baidu Encyclopedia and query log [26]. It has been uniformly proved that the resources help greatly to find subtopic candidates.

Secondly, clustering on subtopic candidates is helpful to find intents which are important for subtopic ranking. For example, ICTIR applies a simple clustering algorithm to group the subtopic candidates [30]. Affinity Propagation algorithm is adopted in HITCSIR system to find intents [26].

We argue that the NTCIR-9 Intent systems can in fact, be further improved using a higher level language unit, say concept or word sense. Our method in this work differs from the previous work by incorporating concept and word sense in subtopic mining and ranking, which is potential particularly to improve recall.

2.2 Intent Ranking

Most NTCIR-9 intent systems rank intents and subtopic based merely on relevance score [28, 26]. MMR model is used in MSINT system to re-rank the subtopics [15]. We do not doubt the contribution of relevance score. But we believe diversity also plays a vital role in intent ranking. We thus propose to measure non-overlapping ratio between intents and incorporate diversity in a unified model for intent ranking.

2.3 Search results clustering

Search result clustering (SRC) aims to facilitate information search. Rather than the results of a query being presented as a flat list, they are grouped on the basis of their similarity and subsequently shown to the user as a list of clusters. Each cluster is intended to represent a different meaning of the input query.

Approaches to search result clustering can be classified as data-centric and description-centric [7]. The former focuses more on the problem of data clustering, while the latter focuses more on the description to produce for each cluster of search results.

A pioneering example of data-centric approaches is Scatter/Gather [11]. The system divides the dataset into a small number of clusters, and performs clustering again and proceeds iteratively after the selection of a group. Developments of this approach have been proposed that improve on cluster quality and retrieval performance [18]. Other data-centric approaches use agglomerative hierarchical clustering [20], exploit link information [31]], and rough sets [21].

Among the most popular and successful description-centric approaches are those based on suffix trees. In order to overcome the low scalability of STC, later developments improved the performance using document-document similarity scores [5]. Crabtree et al. (2005) proposed the Extended Suffix Tree Clustering algorithm (ESTC) with a novel scoring function and a new procedure for selecting the top k clusters to be returned [10]. More recent approaches extract relevant key phrases from generalized suffix trees [3]. Other description-centric approaches are based on format concept analysis [8], single value decomposition [22], link analysis [14], spectral geometry [9], spectral geometry [19] and graph connectivity measures [12]. SRC has also been viewed as a supervised salient phrase ranking task [29].

In this work, we perform search result clustering with the data-centric manner in order to find subtopics of the query.

3. SYSTEM OVERVIEW

3.1 Motivation

We can safely assume that user always carry a definite intent when a query is selected to feed the search engine, but in some cases the query itself does not give enough context to distinguish the intent. We believe that intents underlying a query can be disclosed with context in varying scope. The query itself provides a small context. For example, when the query is "bank account", we know the user is keen on information about finance rather than river. But according the query itself, we still cannot figure out whether he/she is keen on opening a bank account or canceling one. So we need to perform further study within bigger context. Three types of bigger context are worth mentioning. The first one is general knowledge base such as Wikipedia. For the above example, "bank account" can be classified into saving account, current account, etc. by Wikipedia. Some intent might be hidden within Wikipedia knowledge base. The second type of context is query log, which collects a great number of queries input by other users to achieve similar intent. For example, the query "opening bank account" is rather useful to analyze intent underlying the query "bank account". Finally, search engines have achieved a high recall in top 100 results. The true intent might be hidden within the top 100 search results.

The motivation of this work is to understand a specific query with the above four types of context: the query, knowledge base, search results and user behavior statistics on the query. We believe context within the four types can provide very high-recall of subtopic candidates. Precision is another important performance criterion. We can further discover intents by clustering the subtopic candidates, and design a unified model to rank the intents and subtopics based on both relevance and diversity.

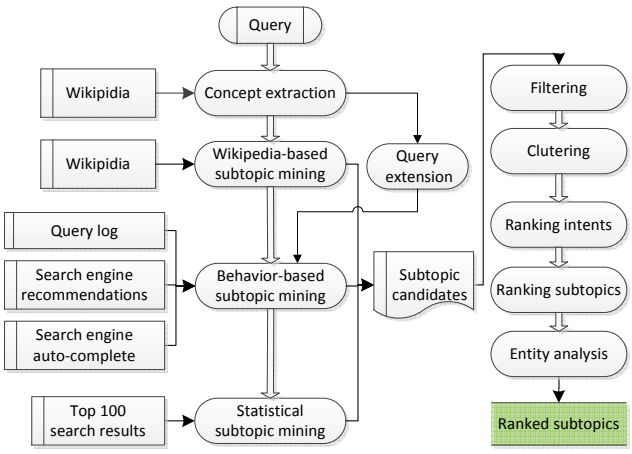

Figure 1: Architecture of THCIB/THUIS intent mining system.

3.2 Architecture

Figure 1 presents architecture of the intent mining system, which is comprised of two modules:

- Subtopic candidate mining module(SCM)

  The SCM module aims at mining subtopic candidates as many as possible from Wikipedia, query log, search engine recommendations, search engine auto-completions and search results.

- Subtopic candidate ranking module(SCR)

  The SCR module aims at precisely ranking relevant subtopic candidates according to both relevance and diversity. In this system, diversity is reflected by the ranked intents discovered in the subtopic candidates with clustering algorithm.

4. SUBTOPIC CANDIDATE MINING

4.1 Extracting Concept(s) from Query

Wikipedia English version provides general knowledge about almost everything. It is a natural source for concept extraction from text [16]. The procedure of concept extraction is given below.

First of all, word(s) within the query are stemmed and tokenized. For example, two concepts are included in the query "battles in the civil war": battle and civil war. Note that we do not remove stop words at this stage as some are included in Wikipedia entry string. We then search within Wikipedia with the original form words, and obtain some Wikipedia entries. We call the entries concepts.

For Chinese queries, we utilize the ICTCLAS[1] to segment the query string into words, and then identify concepts using the Chinese version of Wikipedia.

4.2 Extending the Query

The same intent can be expressed with different queries. We adopt two manners to extend the query.

Firstly, we make use of Wikipedia redirect and disambiguation pages to involve conceptually identical word/phrases. In this way, the initial query is extended to a group of synonymous queries, which are vital for subtopic mining.

Secondly, we propose to construct intent schemas to extract more subtopic candidates. An intent schema includes all concepts in the query, prepositions, and wildcard(s). Prepositions are included since they are indicators of facets. For example, given the query "hobby stores", we construct intent schemas "* of hobby stores", "hobby stores in *", and so on. To find more extensions, we also revise query by adjusting order of the concepts in the query. Search engines give a great number of extra auto-completions and recommendations with the extended queries, which are deemed as important as those provided for the initial query.

In the following step, the initial query and its extensions are all used to discover subtopics from various resources.

4.3 Mining Subtopics in Wikipedia

Wikipedia provides four functions to facilitate subtopic candidate mining: (1) disambiguation, (2) redirects and (3) concept definitions [4]. Besides, we extract related entries which contain the query as a substring.

For every conceptually ambiguous entry, Wikipedia provides a disambiguation page to list all the concepts that the word may carry. For example, the entry for battle can represent surname, military confliction, music, film, and so on, which are listed in the disambiguation page. By parsing the disambiguation page, one is able to find all related concepts, which are helpful to find subtopic candidates.

Wikipedia also provides redirects for an alternative concept. An example of a redirect is to redirect shortest path to shortest path problem. This function is very useful to find synonym of word in a query.

A majority of Wikipedia content is concept definition, which describe various aspects of the concepts. On every content page, there is always a table of content, which provides plentiful subtopic candidates. For example, the query "rock art" can be found in Wikipedia. From the content page, we are able to produce the following subtopic candidates: rock art terminology, rock art background, rock art type, rock art studies, etc. Related entries are those contain the input query as a substring. Taking "rock" for example, there are entries named "rock music", "rock band", and so on, which are all collected as subtopic candidates.

Wikipedia is finely compiled. We use the English version and Chinese version of Wikipedia in THCIB and THUIS respectively. In our experiments, a large proportion of subtopic candidates are discovered within Wikipedia.

4.4 Mining Subtopics in User Behavior Data

What other users input to deliver the similar intent is rather promising in learning intent of the current query. Query log collects a great deal of search behavior data via the search engines. In this work, three kinds of user behavior data are explored [25, 28, 30].

SogouQ provides s huge volume of Chinese query log. To facilitate searching within the query log, we first index SogouQ query log with Lucene. For every query, we search with Lucene with the query string, and obtain a few relevant queries, which are formulated by other users. We use merely the top 10 results as subtopic candidates. For English, there is no such open query log, so we use Anchor Text Query Log for ClueWeb09[2], a simulated query log construct-

[1]http://ictclas.org/

[2]http://lemurproject.org/clueweb09/anchortext-querylog/

ed from anchor text, instead. And to get better results, we search the English query with the original form words.

The remaining two types of user behavior data are actually popular products of search engines, i.e., search recommendations and auto-completions. For example, when we input "battles in the civil war" into Google search engine, we obtain the following auto-completes: battles in the civil war timeline, how many battles in the civil war, important battles in the civil war, who won more battles in the civil war, etc. After we click the search button, we can find a few search recommendations at the bottom of the search results page: after the civil war, us civil war, English civil war, Chinese civil war, etc. We believe intents are probably hidden in the user behavior data. NTCIR10 Intent Mining task organizer provides search recommendations and auto-completions. So we used all the recommendations and auto-completes as subtopic candidates.

4.5 Mining Subtopics in Search Results

Previous work performs clustering on search results to provide a friendly search results interface. Search result clustering is promising in discovering intents within the search results [15].

We adopted the word sense induction (WSI) framework to discover conceptual aspects of the query within the search results (Google for English and Sogou for Chinese). To alleviate complexity, we only use the title and snippets of the top 100 results that contain the query completely or partially. We use the Bayesian model [6] to induce word senses which outperforms the state-of-the-art systems in SemEval-2007 evaluation [1]. In this model, LDA is used in the contexts of each query which referred to search results of each query in this paper. After LDA, we take the top word of each topic together with the query itself as intent.

Inspired by the ISCAS system in NTCIR-9 [17], we also collect the titles of the top 100 search results as subtopic candidates.

4.6 Subtopic Candidate Filtering

For some queries, we obtained more than 1000 subtopic candidates, which are sufficient for intent mining. Before we perform subtopic ranking, we assign the following rules to exclude the less likely subtopic candidates:

Rule #1: Candidates that are contained in the query are excluded.

Rule #2: Candidates that do not contain all concepts (or corresponding synonymous concepts) of the query are excluded.

In rule #2, concepts are extracted from the query with Wikipedia knowledge base (see details in Section 4.1 and 4.2)

In our experiments, 28.8 percent of subtopic candidates are deleted after filtering. The filtering procedure reduces not only complexity but also noise in subtopic ranking.

5. SUBTOPIC RANKING

We argue that rank of a subtopic is determined by three factors: relevance of the subtopic, importance of the subtopic source, and significance of the intent that the subtopic belongs to. The final ranking of a subtopic t is determimed by $w_{ST}(t)$, $w_{SC}(t)$ and $w_{IN}(t)$, where $w_{ST}(t)$ denotes relevance score of the subtopic, $w_{SC}(t)$ importance score of the source that the subtopic comes from, and $w_{IN}(t)$ significance score of the intent where the subtopic belongs to.

Relevance of a subtopic is assigned the frequency of the initial query within the top 1,000 search results requested with the subtopic as a query. As for the importance score of the subtopic source, the popular method is assigning empirical weights to the involved sources. We follow this method in our experiments.

Calculating significance score of a hidden intent is more complicated. We first perform clustering on the subtopic candidates to discover intents. We also perform knowledge based entity analysis on subtopic candidates to resolve homogenous entities so as to refine the intents. More details are presented in 5.1, 5.2 and 5.3.

The ranking procedure can be summarized as follows.

1. Rank the subtopic candidates in the declining order of $w_{ST}(t) + w_{SC}(t)$.
2. First calculate the $w_{IN}(t)$ of each cluster after clustering, rank the intents in declining order, sort the subtopic in each cluster in descending order by $w_{ST}(t)+ w_{SC}(t)$. Then, iteratively get the top subtopic candidate in each cluster until all subtopic candidates are returned.
3. After we get a ranked list of subtopics, we apply entity analysis and enlarge the distance of homogenous subtopics to enhance diversity.

5.1 Subtopic Clustering

Relevance is important in subtopic mining, but it is not the only thing. Another important issue in subtopic ranking is diversity. Our ranking roadmap is first discovering exclusive intents from the subtopic candidates and the intents as well as relevance scores are in turn used to rank the subtopic candidates. We apply clustering algorithm to organize the subtopic candidates into a few clusters [28], which is referred to as implicit intents. As the purpose is ranking, no labeling job is involved.

In THCIB and THUIS system, we apply Affinity propagation as the clustering algorithm. Affinity propagation (AP) is a clustering algorithm that has been introduced by Frey and Dueck (2009) [13]. The AP algorithm has been applied in various fields. Exemplars are identified among data points, and clusters of data points are formed around these exemplars. It operates by simultaneously considering all data point as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges.

AP clustering algorithm has two initial inputs: a similarity matrix M where M_{ij} represent how points i prefers point j to be the exemplar and a preference list which represents how likely the point should be an exemplar.

In our experiments, we use the sense-based similarity between subtopic candidates as the input matrix. The similarity is calculated in the following steps. First we apply Word Sense Induction [6] to get different senses of each subtopic candidate. Then, we choose the most similar senses between each pairs of subtopic candidates and calculate their cosine similarity.

As for the preference input, we have two versions.

- Standard AP algorithm

  We use the mean value of the similarity matrix as the input preference value for all points.

- Revised AP algorithm

  As our subtopic candidates are collected from different sources and have their own relevance score, which shows the importance of each subtopic candidate, the preference of subtopic candidate t is calculated as follows.

$$p_t = w_{ST}(t) + w_{SC}(t)$$

5.2 Entity Analysis

We encounter a large proportion of named entities for person, organization and location in the subtopic candidates. Consider two subtopic candidates: furniture for small spaces New York, furniture for small spaces Los Angeles. Obviously, they refer to furniture for small spaces in two cities. We call entities like New York and Los Angeles exclusive entities.

Exclusive entities sometimes lead to intent fission. Entity analysis helps to detect fission of this type so that we can restore the appropriate intent from the fissions. For such cases, we adopt Freebase [3] to recognize the entities and generalize subtopic candidates with Freebase so as to associate named entities with the same ontology type to some ontological clusters.

Freebase is an online collection of structured data, aiming to create a global resource which allows people (and machines) to access common information more effectively. Under certain license, Freebase provides a JSON based HTTP API for programmers to develop applications on any platform to utilize the Freebase data.

Since no Chinese version is proved by Freebase, there is no such module in our THUIS system.

5.3 Ranking the Intents

We propose that the intents are ranked based on relevance score of the subtopic candidates and important score of the source. Considering certain intent containing N subtopic candidates $\{t_i\}(i = 1...N)$, significance score of the intent is calculated as follows.

$$w_{IN} = \sum_{i=1}^{N} [w_{ST}(t_i) + w_{SC}(t_i)]$$

where $w_{ST}(t)$ denotes relevance score of the subtopic, and $w_{SC}(t)$ importance score of the source.

Not all discovered intents are considered in subtopic ranking. Considering relevance and diversity at the same time, we propose to consider the intents that include more than 5 subtopic candidates.

6. EVALUATION

6.1 Evaluation Metric

Three widely-used metrics are used:

- D-nDCG which measures overall relevance across intents [24];
- I-rec which measures diversity [2].
- D#-nDCG is a linear combination of I-rec and D-nDCG [25].

For the metrics, performance scores in top 10, 20 and 30 results are evaluated.

[3]http://www.freebase.com/

6.2 Submitted Runs

THCIB has submitted the following five runs for the English Subtopic Mining subtask:

- THCIB-S-E-1A: (1) explores search recommendations (provided by NTCIR10), search auto-completions (provided by NTCIR10), related webpages (Google), query log (ClueWeb09) and semantic descriptions (Wikipedia) to obtain concept-level subtopic candidates of each query; (2) ranks the subtopic candidates according to source weights and word frequencies in search result snippets.
- THCIB-S-E-2A: (1) obtains subtopic candidates with THCIB-S-E-1A system; (2) generates extended queries and get the user behavior statistics from search engine (see details in 4.2); (3) ranks the subtopic candidates according to source weights and word frequencies in search result snippets.
- THCIB-S-E-3A: (1) obtains subtopic candidates with THCIB-S-E-2A system; (2) generalizes subtopic candidates with Freebase so as to associate named entities with the same ontology type to some ontological clusters; (3) ranks the subtopic candidates according to source weights, ontological clusters and word frequencies in search result snippets.
- THCIB-S-E-4A: (1) executes step (1) and (2) in THCIB-S-E-3A system. (2) clusters subtopic candidates based on semantic similarity with standard AP algorithm; (3) ranks the subtopic candidates according to source weights, ontological clusters, semantic clusters and word frequencies in search result snippets.
- THCIB-S-E-5A: revises THCIB-S-E-4A system by replacing the standard AP algorithm with a revised AP algorithm (see details in Section 5.1).

THUIS has submitted the following four runs for the Chinese Subtopic Mining subtask:

- THUIS-S-C-1A: (1) explores search recommendations (provided by NTCIR10), related webpages (Sogou), query log (SogouQ) and semantic descriptions (Wikipedia) to obtain concept-level subtopic candidates of each query; (2) ranks the subtopic candidates according to source weights and word frequencies in search result snippets.
- THUIS-S-C-2A: see THCIB-S-E-2A.
- THUIS-S-C-3A: see THCIB-S-E-4A without entity analysis.
- THUIS-S-C-4A: see THCIB-S-E-5A without entity analysis.

We submit THUIS-S-C-1A to evaluate our baseline system compared with the THUIR system, one of the best systems in NTCIR-9. The other runs are submitted for the aim of verifying the effect of query extension, entity analysis, standard AP and revised AP in sequence. The submissions for Chinese are the same.

6.3 Resources and Weights

Multiple resources are used in THCIB and THUIS system. The resources as well as their weights are presented in Table 1 and Table 2.

Table 1: Resource and their weights for THCIB

Source name	Weight
Bing Completion	1
Bing Suggestion	1
Google Completion	1
Yahoo Completion	1
Query extension	0.9
Query Log	0.2
SRC	0.4
Search Result Title	0.2
Wiki Concept Definition	0.8
Wiki Disambiguation & Redirects	1
Wiki Related Entries	0.8

Table 2: Resource and their weights for THUIS

Source name	Weight
Bing Suggestion	1
Baidu Suggestion	1
Sogou Suggestion	1
Google Suggestion	1
Query extension	0.9
Query Log	0.6
SRC	0.4
Search Result Title	0.2
Wiki Concept Definition	0.8
Wiki Disambiguation & Redirects	1
Wiki Related Entries	0.8

6.4 Results

Table 3 presents the official SM evaluation results of five runs of THCIB system for the English subtopic mining. Table 4 presents the official SM evaluation results of four runs of THUIS system for the Chinese subtopic mining[27]. Note the results on metrics of top 20 and top 30 are calculated by ourselves with the gold standard published by the organiser.

6.5 Discussion

We observe how the systems perform on the overall dataset as well as the individual topics.

System Analysis

According to Table 3, THCIB-S-E-2A outperforms THCIB-S-E-1A. We can conclude concept-based query expansion helps to recall more relevant subtopic.

Comparing THCIB-S-E-4A and THCIB-S-E-5A, we find the revised AP algorithm outperforms the standard AP algorithms in most evaluation metrics.

Comparing THCIB-S-E-2A, THCIB-S-E-3A and THCIB-S-E-5A, we find the unified ranking model do not bring performance gain on this dataset. There are 2 reasons that may explain it. First, the strategy that we use clustering and entity analysis in intent ranking is relatively simple, so the two modules may not work as well as expected. Second, each topic has no more than 9 intents in the gold standard, which may have a little negative effect on our system. So we believe the unified ranking model, though not outperforms the state-of-the-art system, is potential to improve. We will focus more on how to use clustering and entity analysis technology to rank intents in the future.

Table 3: Evaluation results of English Subtopic Mining runs

cut-off	run name	I-rec	D-nDCG	D#-nDCG
@10	THCIB-S-E-1A	0.3785	0.3384	0.3584
	THCIB-S-E-2A	**0.3797**	**0.3499**	**0.3648**
	THCIB-S-E-3A	0.3681	0.3383	0.3532
	THCIB-S-E-4A	0.3502	0.3323	0.3413
	THCIB-S-E-5A	0.3662	0.3215	0.3438
@20	THCIB-S-E-1A	0.5769	0.3274	0.4522
	THCIB-S-E-2A	**0.5899**	**0.3406**	**0.4653**
	THCIB-S-E-3A	0.5544	0.3251	0.4397
	THCIB-S-E-4A	0.477	0.2784	0.3777
	THCIB-S-E-5A	0.5395	0.304	0.4218
@30	THCIB-S-E-1A	**0.693**	0.3177	**0.5054**
	THCIB-S-E-2A	0.6743	**0.3284**	0.5014
	THCIB-S-E-3A	0.6486	0.3244	0.4865
	THCIB-S-E-4A	0.5855	0.2691	0.4273
	THCIB-S-E-5A	0.6339	0.2986	0.4662

Table 4: Evaluation results of Chinese Subtopic Mining runs

cut-off	run name	I-rec	D-nDCG	D#-nDCG
@10	THUIS-S-C-1A	0.3381	**0.4923**	**0.4402**
	THUIS-S-C-2A	0.3622	0.4157	0.389
	THUIS-S-C-3A	0.3953	0.4504	0.4228
	THUIS-S-C-4A	**0.4036**	0.462	0.4328
@20	THUIS-S-C-1A	**0.5322**	**0.4776**	**0.5049**
	THUIS-S-C-2A	0.4467	0.3385	0.3926
	THUIS-S-C-3A	0.5067	0.3969	0.4518
	THUIS-S-C-4A	0.5163	0.4215	0.4689
@30	THUIS-S-C-1A	**0.5842**	**0.4677**	**0.5259**
	THUIS-S-C-2A	0.5249	0.3272	0.426
	THUIS-S-C-3A	0.5571	0.3814	0.4692
	THUIS-S-C-4A	0.5636	0.3764	0.47

Similar observations are made on the Chinese task except that query extension lead to performance degradation, which may be caused by the difference in language features and search engines between Chinese and English.

Per-Topic Analysis

We then observe how systems perform on different topics. Due to space limit, the observation is conducted only on English queries, and only D#-nDCG values of the five systems on the 50 queries are presented in Figure 2.

Seen from Figure 2, the systems perform differently upon the topics. To be more specific, system THCIB-S-E-1A yields 8 best, THCIB-S-E-2A 13 best, THCIB-S-E-3A 6 best, THCIB-S-E-4A 13 best, and THCIB-S-E-5A 10 best. This indicates that there is actually no system which is consistently advantageous over the others. We thus conclude that the algorithms used in the systems are equally interesting. There arises a natural question what type of queries the systems can handle better. We currently have no answer, which will be studied in our future work.

We further study relationship between system performance and query length (i.e., number of word the query contains). We insert the curve representing query length of each topic and produce Figure 3.

Seen from Figure 3, there is no intuitive relation between system performance and query length. It can thus be concluded that performance of our system is not sensitive to length of the query, but other issues. Due to time limit, we have not conduct further analysis on nature of the queries. For example, queries in different domain might be one issue. Another issue could be ambiguity nature. Future work will be conducted to discover these issues.

7. CONCLUSION

Understanding the query is a challenging issue to information retrieval. In this work, we propose to incorporate concept and word sense in subtopic mining and ranking, which brings marginal performance gain. We also find that

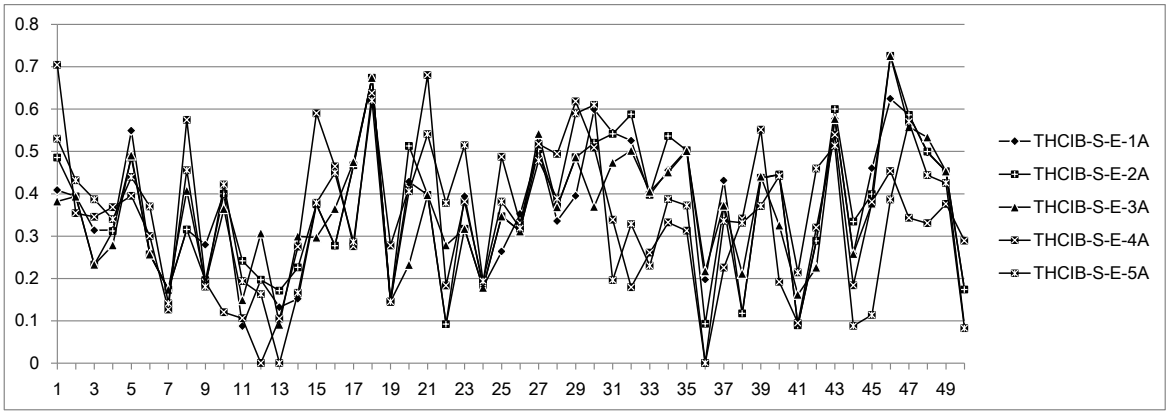

Figure 2: System performance upon topics.

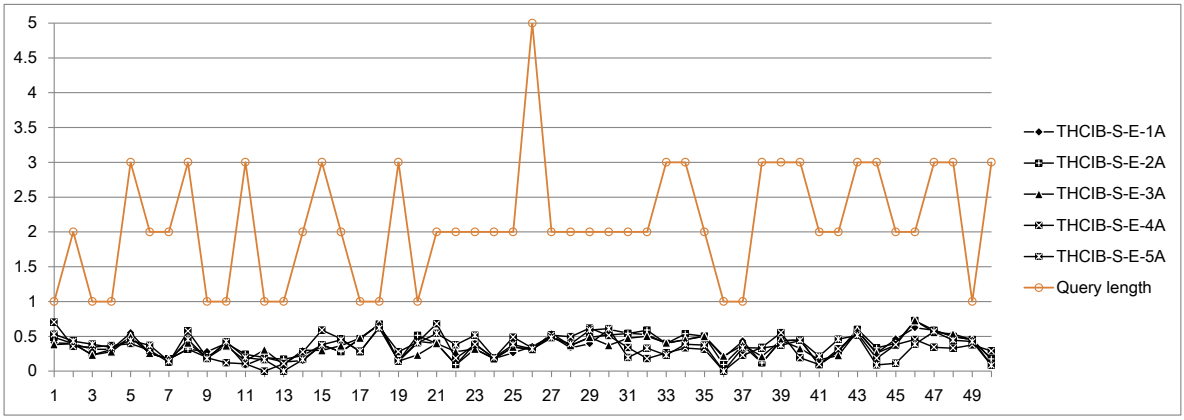

Figure 3: System performance regarding query length.

the unified intent ranking model is promising in producing satisfactory results. Experimental results also disclose that subtopic clustering and named entity analysis fail to improve performance. But we think it is still early to make the negative conclusion as experiments in this work are preliminary due to time limit. To explore the reasons, future work is planned to tune the system parameters.

ACKNOWLEDGEMENT

This work is supported by Canon Inc. (No. TEMA2012).

8. REFERENCES

[1] E. Agirre and A. Soroa. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, 2007.

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

[3] A. Bernardini, C. Carpineto, and M. D'Amico. Full-subtopic retrieval with keyphrase-based search results clustering. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 206–213. IET, 2009.

[4] C. Bøhn and K. Nørvåg. Extracting named entities and synonyms from wikipedia. In *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, pages 1300–1307. IEEE, 2010.

[5] S. Branson and A. Greenberg. Clustering web search results using suffix tree methods. Technical report, Stanford University, Tech. Rep. CS276A Final Project, 2002.

[6] S. Brody and M. Lapata. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111. Association for Computational Linguistics, 2009.

[7] C. Carpineto, S. Osiński, G. Romano, and D. Weiss. A survey of web clustering engines. *ACM Computing Surveys (CSUR)*, 41(3):17, 2009.

[8] C. Carpineto and G. Romano. Exploiting the potential of concept lattices for information retrieval with credo. *Journal of universal computer science*, 10(8):985–1013, 2004.

[9] D. Cheng, R. Kannan, S. Vempala, and G. Wang. A divide-and-merge methodology for clustering. *ACM*

Transactions on Database Systems (TODS), 31(4):1499–1525, 2006.

[10] D. Crabtree, X. Gao, and P. Andreae. Improving web clustering by cluster selection. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 172–178. IEEE, 2005.

[11] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 318–329. ACM, 1992.

[12] E. Di Giacomo, W. Didimo, L. Grilli, and G. Liotta. Graph visualization techniques for web clustering engines. *Visualization and Computer Graphics, IEEE Transactions on*, 13(2):294–304, 2007.

[13] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[14] F. Gelgi, H. Davulcu, and S. Vadrevu. Term ranking for clustering web search results. In *Proc. of WebDB*, volume 7. Citeseer, 2007.

[15] J. Han, Q. Wang, N. Orii, Z. Dou, T. Sakai, and R. Song. Microsoft research asia at the ntcir-9 intent task. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 116–122, 2011.

[16] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 179–186. ACM, 2008.

[17] X. Jiang, X. Han, and L. Sun. Iscas at subtopic mining task in ntcir9. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 168–171, 2011.

[18] W. Ke, C. R. Sugimoto, and J. Mostafa. Dynamicity vs. effectiveness: studying online clustering for scatter/gather. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM, 2009.

[19] Y. Liu, W. Li, Y. Lin, and L. Jing. Spectral geometry for simultaneously clustering and ranking query search results. In *Proc. of SIGIR*, volume 8, pages 539–546, 2008.

[20] Y. S. Maarek, R. Fagin, I. Z. Ben-Shaul, and D. Pelleg. Ephemeral document clustering for web applications. 2000.

[21] C. L. Ngo and H. S. Nguyen. A method of web search result clustering based on rough sets. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 673–679. IEEE, 2005.

[22] S. Osinski and D. Weiss. A concept-driven algorithm for clustering search results. *Intelligent Systems, IEEE*, 20(3):48–54, 2005.

[23] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM, 2004.

[24] T. Sakai and R. Song. 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*, pages 1043–1052. ACM, 2011.

[25] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the ntcir-9 intent task. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 82–105, 2011.

[26] W. Song, Y. Zhang, H. Gao, T. Liu, and S. Li. Hitscir system in ntcir-9 subtopic mining task. In *Proceedings of NTCIR-9 Workshop Meeting*, 2011.

[27] S. Tetsuya, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the ntcir-10 intent-2 task. In *Proceedings of NTCIR-10 Workshop Meeting*, 2013.

[28] Y. Xue, F. Chen, T. Zhu, C. Wang, Z. Li, Y. Liu, M. Zhang, Y. Jin, and S. Ma. Thuir at ntcir-9 intent task. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 123–128, 2011.

[29] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217. ACM, 2004.

[30] S. Zhang, K. Lu, and B. Wang. Ictir subtopic mining system at ntcir-9 intent task. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 106–110, 2011.

[31] X. Zhang, X. Hu, and X. Zhou. A comparative evaluation of different link types on enhancing document clustering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 555–562. ACM, 2008.