ABSTRACT
In this report, we describe our experimental results for the NTCIR-9 INTENT task. For our experiments, we use our experimental search engine, NewT. NewT is a ranked self-index capable of supporting multiple languages by deferring linguistic decisions until query time. To our knowledge, this is the first Information Retrieval task on the ClueWeb09-JA collection performed entirely with ranked self-indexes.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—indexing methods; H.3.2 [Information Storage and Retrieval]: Information Storage—file organization; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—query formulation, retrieval models, search process; I.7.3 [Document and Text Processing]: Text Processing—index generation

General Terms
Text Indexing, Text Compression, Language Independent Text Indexing, Data Storage Representations, Experimentation, Measurement, Performance

1. INTRODUCTION
In this paper, we describe our experimental approach for the NTCIR-9 INTENT task. We present a novel approach to language independent, ranked document retrieval using an enhanced version of the self-indexing algorithms used in NewT. NewT is a ranked self-index capable of supporting multiple languages by deferring linguistic decisions until query time. Our search engine uses backwards search in a Burrows-Wheeler transform similar to the FM-index of Ferragina and Manzini [6] and a wavelet tree [7] for occurrence counting in a document array [13]. To our knowledge, this is the first Information Retrieval task on the ClueWeb09-JA collection performed entirely with ranked self-indexing algorithms.

For this year’s task, our group only participated in the Japanese subtasks as we did not have a Chinese collaborator to assist with language specific issues. Our primary goal for NTCIR-9 was to implement and test our new class of indexing algorithms for multilingual tasks. Our initial evaluation of Japanese queries led to an unexpected problem with character-level indexing, overlapping substrings.

2. BACKGROUND
The CJKV language family is traditionally indexed as character n-grams or terms. Since Japanese is an unsegmented language, either of these approaches is possible, but recent approaches have focused primarily on morphological term parsing. Conversely, recent work on Chinese tokenization has used character n-grams [10, 15, 11]. Generally, character or n-gram based methods increase recall, while word-based methods improve precision [11].

The word segmentation process is generally performed on Japanese documents by using morphological analyzers, such as ChaSen1 and MeCab2. However, the output of these morphological analyzers is not always correct. Some morphemes suggested by morphological analyzers are overly segmented, especially when documents contain unknown words [22]. Furthermore, output texts are under-segmented or not segmented at all when web pages omit punctuation marks or white space in the text for the sake of simplicity or layout. When wrongly segmented morphemes are stored as index terms, the search results can be poor.

Either of the CJKV methods depend on an inverted index for term or n-gram indexing. Inverted indexes are the dominant solution for IR search [24]. However, forcing CJKV languages to rely on morphological analysis is not always the best option. Character-based methods have certain advantages over term-based methods, but the additional space overhead makes these methods infeasible for large collection. Recently, self-indexing algorithms have received a great deal of attention because of their efficient search capabilities and reduced space overhead [14]. These indexing algorithms have interesting theoretical and practical performance on basic pattern matching operations, but ranked search capabilities on large datasets is still open [8].

In this report, we investigate the problem of using self-indexing algorithms to solve the ranked document search problem. In order to solve the document search problem, basic self-indexing algorithms are not sufficient. However, using an auxiliary data structure to manage a document array enables basic document ranking [13, 19, 4]. For the INTENT task, we use an enhanced version of the greedy top-k approach described in [4]. A full discussion of prior work on applications of self-indexes to the ranked document search problem is beyond the scope of this report. Please refer to [4] or [5] for a more comprehensive discussion of the algorithms used in NewT.

1http://chasen-legacy.sourceforge.jp/
2http://mecab.sourceforge.net/
All prior published work on ranked self-indexes use a trivial TF×IDF ranking metric, and have generally focused on phrase queries instead of bag-of-words queries. For the INTENT task, two bag-of-words ranking functions were implemented. The first metric is referred to as raw term frequency ranking. For this metric, we simply compute the aggregate of raw frequency counts per document, \( f_{t,d} \), for each term or substring, \( t \).

\[
\text{RAW} = \sum_{t \in q} f_{t,d}
\]

However, the RAW ranking metric does not normalize for document length or \( \text{idf} \). Therefore, we also implemented a simple BM25 variant as follows:

\[
\text{BM25} = \sum_{t \in q} \log \left( \frac{N - f_t + 0.5}{f_t + 0.5} \right) \cdot \text{TF}_{\text{BM25}}
\]

\[
\text{TF}_{\text{BM25}} = \frac{f_{t,d} \cdot (k_1 + 1)}{f_{t,d} + k_1 \cdot ((1 - b) + b \cdot \ell_d / \ell_{avg})}
\]

Here, \( N \) is the number of documents in the collection, \( f_t \) is the number distinct documents appearances of \( t \), \( k_1 = 1.2 \), \( b = 0.75 \), \( \ell_d \) is the number of UTF8 symbols in the documents, and \( \ell_{avg} \) is the average of \( \ell_d \) over the whole collection. For self-indexes, there is an efficiency trade-off between locating the top-\( k \) \( f_{t,d} \) values and accurately determining \( f_t \) since the index can extract exactly \( k \) \( f_{t,d} \) values without processing every document. We leave the task of finding the most efficient trade-off between these competing requirements, and devising ranking metrics that do not make term-based \textit{independent} and identically distributed assumptions as a topic of future work.

### 2.1 Diversity Ranking

User satisfaction in web search has received significant attention in recent years. While there may be several possible “senses” for a search topic, a user typically wishes to find only one meaning for the topic. In such a situation, browsing many unrelated documents in the ranked list quickly leads to user dissatisfaction. To solve this problem, the Maximal Marginal Relevance (MMR) ranking method minimizes redundancy in the search results. Since diversification was recognized as a challenging problem, experiments on diversity ranking have been conducted in the TREC diversity track [2, 3].

The NTCIR-9 INTENT task [18] focuses on the diversity ranking problem in a manner similar to the TREC Web track, but with a focus on Japanese and Chinese linguistic disambiguation. In the NTCIR-9 INTENT task, subtopic strings were not provided by the organizer, but participants were asked to prepare subtopic strings and submit them in the Subtopic Mining subtask. In the TREC diversity task, subtopics were extracted from the logs of a commercial search engine [2, 3]. Prior to the TREC task, clustering of similar queries [23, 21] and understanding query-query reformulations by users [16] were explored to obtain subtopics from a query log. Unfortunately, a query log or other additional resources were not provided in the Japanese Subtopic Mining subtask of the NTCIR-9 INTENT task.

The diversity task of the TREC Web track [2, 3] introduced two topic types: faceted and ambiguous. Faceted topics are topics that are not ambiguous by themselves, but have different related subtopic strings. Ambiguous topics are topics that are polysemous, and need to be clarified with subtopic strings. To deal with ambiguous queries, query expansion using lexical-semantic relations was studied [20]. Lexical resource such as WordNet\(^3\) and Wikipedia\(^4\) have been used for resolving word sense disambiguity [12] and for named entity recognition [9]. For the INTENT task, we used the Japanese edition of Wikipedia to obtain subtopic strings for both faceted and ambiguous topics.

### 3. EXPERIMENTAL FRAMEWORK

In this section, we describe the experimental setup used for the NTCIR-9 INTENT task.

#### 3.1 Collection Processing and Indexing

For the INTENT task, we indexed the ClueWeb09-JA collection in two steps. First, each document from the collection was extracted and normalized using Lynx,\(^5\) a text-based browser with an option to output processed and formatted texts. The extracted documents are converted to UTF8's character code. Next, all whitespace was removed from each document to create a contiguous UTF8 string, followed by a distinct end of document identifier. The fully processed ClueWeb09-JA collection was partitioned into blocks of 500,000 documents and indexed with Newt. Since Newt is an in-memory index, each 500k document block is serialized to disk and processed separately at query time.

#### 3.2 Topic Processing

In order to evaluate Newt, we processed the search topics and generated two vanilla query sets (Vanilla 1, Vanilla 2) and a diversity query set (Diversity 1). Table 1 has a brief descriptions for each of our query sets. In the table, attributes in parenthesis represent the corresponding runs in Table 3.

#### 3.2.1 Vanilla Queries

For the two vanilla query sets, word segmentation of search topics was performed using MeCab. Hence, search topics were decomposed into morphemes, or shorter sub-strings of the search topics. This is a standard approach for query processing in term-based document retrieval in Japanese text. Hence, we refer to these query sets as

\[^3\]http://wordnet.princeton.edu/
\[^4\]http://www.wikipedia.org/
\[^5\]http://lynx.isc.org/

<table>
<thead>
<tr>
<th>Query Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla 1</td>
<td>Word segmentation of original search topics using MeCab, retaining all morphemes. (RMIT-D-J-4)</td>
</tr>
<tr>
<td>Vanilla 2</td>
<td>Word segmentation of original search topics using MeCab, retaining only nouns. (RMIT-D-J-2, RMIT-D-J-5)</td>
</tr>
<tr>
<td>Diversity 1</td>
<td>Original search topics without word segmentation, with addition of subtopic strings. (RMIT-D-J-1, RMIT-D-J-3)</td>
</tr>
</tbody>
</table>

Table 1: Summary of query sets.
“vanilla”. Specific character strings in search topics are resolved into common words in relevant documents using morphological analysis. However, the morphological analysis may also produce ambiguous or irrelevant character substrings from search topics, resulting in performance degradation.

Vanilla 1 and Vanilla 2 use a slight variation in query expansion: all morphemes were used in Vanilla 1, but only nouns were used in Vanilla 2, with prefixes, suffixes, and particles being dropped. Vanilla 1 is therefore essentially a simple parse of the original topic, with little or no modification, and we treat Vanilla 1 as a baseline in our analysis.

### 3.2.2 Query Diversification

For the diversity query set, search topics were retained without word segmentation. Hence, queries in this set gather target documents for search topics. Although the queries in this set faithfully represent the original character strings of the search topics, some of them are overly specific, resulting in low recall. At the other extreme, some queries are short and ambiguous, resulting in lower precision. In both cases, query expansion is an effective approach. Subtopic strings that were obtained from the Japanese Wikipedia for each distinct topic sense were used for expansion.

**Topic Disambiguation:** In order to gather distinct subtopic senses, we used “Disambiguation in Wikipedia”, which resolves the conflicts that arise when a single term is ambiguous. Specifically, we used three types of disambiguation definitions from Wikipedia: (1) pages, (2) titles, and (3) “See” in articles. For disambiguation pages, we used the Japanese character string, “曖昧さ回避” (meaning “disambiguation page” in English) as a pattern to identify disambiguation pages for search topics. To identify “Disambiguation in Wikipedia” in titles, we also used a set of round parentheses (“(“ and “)” as tokens for pattern matching. For example, a reader of Wikipedia may find the page “Orange”, which is a disambiguation page because it shows “disambiguation page” in the page top, and lists various meanings of “Orange” and links. In the page, major meanings of “Orange” are titled with round parentheses, “Orange (fruit),” “Orange (colour),” etc. Some ambiguous topics are not consolidated in disambiguation pages. Hence, we also used Perl Compatible Regular Expressions (PCRE), incorporated into the scripting language PHP to gather disambiguation definitions from the headings of articles. Specifically, we used the Japanese character strings and regular expressions “." to obtain disambiguation definitions from articles.

**Wikipedia summaries in topics:** If a search topic had no disambiguation definitions, but had an article, we obtained the last noun from the first sentence in the article. We assume that such topics are fairly specific because no disambiguation exists on Wikipedia, but the descriptions can help to make ambiguous queries more specific. According to Wikipedia’s Style Manual, the introduction of a Wikipedia article is a summary of the most important aspects. Extracting all aspects from the introduction is preferred, but difficult to automate. The automation would require a precise analysis of the dependency structure in Japanese. In this task, we performed word segmentation on the first sentence by using MeCab, and simply chose the last noun from the result of morphological analysis. For example, a reader of Wikipedia may find the sentence “マカロンを使ったフランスを代表する洋菓子である," meaning “A macaroon is a western confectionery, which is made with almond and a representative of France,” as the first sentence in the page “マカロン” meaning “Macaroon.” This sentence contains the five nouns, “マカロン,” “アーモンド,” “代表," “フランス," “洋菓子” meaning “macaroon,” “almond,” “representative," “France,” “western confectionery," respectively. We obtained the last noun in the sentence, which is “洋菓子” meaning “western confectionery” as an important aspect for the topic. The expanded query for the example is “マカロン, 洋菓子” meaning, “macaroon, western confectionery” in English.

**Topics as queries:** Some search topics were not listed in the Japanese Wikipedia. For those topics, no diversification was applied to the queries. However, some topics were excessively specific. Consequently, the queries could not be applied directly to the collection. In these instances, the query was represented as both the search topic and all of the nouns in the search topic. For example, with the excessively specific search topic “スタバのシナモンロールのレシピ,” meaning “cooking recipes for cinnamon rolls in Starbucks stores” in English is used as a query, too few results are returned. So, both the topic and all the nouns are used in the query, “スタバのシナモンロールのレシピ, シナモンロール, シナモン, レシピ” meaning “cooking recipes for cinnamon rolls in Starbucks stores, Starbucks stores, cinnamon rolls, cooking recipes” in English.

### 3.3 Description of Runs

We submitted a single run for topic mining and 5 runs for document ranking. Tables 2 and 3 give brief descriptions of our submitted runs. In Table 3, attributes in parenthesis represent the corresponding query sets in Table 1.

MeCab, a dictionary based morphological analyzer, was used for topic processing. By default, MeCab uses a recommended standard dictionary, the IPA dictionary. Although the IPA dictionary contains a large number of entries for Japanese morphemes, some search topics were not in the vocabulary. In our preliminary experiment, some search topics were tagged as unknown words by the morphological analysis, and meaningful morphemes were segmented into shorter substrings. If morphemes in queries are overly segmented into short strings, search results

<table>
<thead>
<tr>
<th>Run</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMIT-S-J-1</td>
<td>Obtain subtopic strings using pattern matching from Japanese Wikipedia.</td>
</tr>
</tbody>
</table>

Table 2: Subtopic mining run.
dictionaries in the topic processing of the as nouns in our experimental dictionary. We used both as Wikipedia titles. All newly added words were treated on. To reduce the risk of this problem, we leveraged the for “macintosh,” “aromatheraphy,” “onomatopoeia” and so "aro,” “on” , search results are intermingled with documents "macaroon" is segmented into three short strings, “mac,” tend to contain more irrelevant documents, and the search performance is notably degraded. For instance, if the word “macaroon” is segmented into three short strings, “mac,” “aro,” “on”, search results are intermingled with documents for “macintosh,” “aromatherapy,” “onomatopoeia” and so on. To reduce the risk of this problem, we leveraged the Japanese Wikipedia title list to expand the dictionary for MeCab. We defined dictionary entries for search topics separately to a depth of $k_{max}$. Next, a single answer set $A$ is generated by a round robin merge of all of the Wikipedia Sense lists $W_1 \ldots W_s$. For each list $i$, the next highest ranked document in list $W_i$ is found and appended to $A$ to a total length $k_{max}$. This ensures that each sense contributes the same percentage of documents to the final ranked list.

**Algorithm 1 Wikipedia Round Robin Merge Algorithm**

**input:** A collection of $s$ ranked document lists, $W_1 \ldots W_s$, of length $k_{max}$.

**output:** A ranked list of $k_{max}$ documents, $A$.

1: $i \leftarrow 1$
2: $A \leftarrow \{\}$
3: $doc \leftarrow \text{First}(W_i)$
4: for $k = 1$ to $k_{max}$ do
5: \hspace{1em} while $doc \in A$ do
6: \hspace{2em} set $doc \leftarrow \text{Successor}(W_i)$
7: \hspace{1em} end while
8: \hspace{1em} Append($A, doc$)
9: \hspace{1em} set $i \leftarrow (i \mod s) + 1$
10: end for
11: return $A$

<table>
<thead>
<tr>
<th>Run</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMIT-D-J-1</td>
<td>Simple BM25 ranking with Newt using round robin query diversification. (Diversity 1)</td>
</tr>
<tr>
<td>RMIT-D-J-2</td>
<td>Simple BM25 ranking with Newt using naive query expansion. (Vanilla 2)</td>
</tr>
<tr>
<td>RMIT-D-J-3</td>
<td>Simple TF ranking with Newt using round robin query diversification. (Diversity 1)</td>
</tr>
<tr>
<td>RMIT-D-J-4</td>
<td>Simple BM25 ranking with Newt using unmodified topic queries. (Vanilla 1)</td>
</tr>
<tr>
<td>RMIT-D-J-5</td>
<td>Simple TF ranking with Newt using naive query expansion. (Vanilla 2)</td>
</tr>
</tbody>
</table>

**Table 3: Document ranking runs.**

<table>
<thead>
<tr>
<th>Run</th>
<th>Vanilla 1</th>
<th>Vanilla 2</th>
<th>Diversity 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMIT-D-J-1</td>
<td></td>
<td></td>
<td>★ Pattern Match</td>
</tr>
<tr>
<td>RMIT-D-J-2</td>
<td>★ BM25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMIT-D-J-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMIT-D-J-4</td>
<td>★ BM25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMIT-D-J-5</td>
<td>★ TF</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4: Run and query set correspondence.**

In this section we present the results of our experiments for the Intent task.

### 4.1 Evaluation Metrics

The official effectiveness performance measures for the intent subtopic mining and document ranking tasks are: $I$-rec, which measures the proportion of intents covered by the documents in the search results list; D-nDCG, which uses a global gain to measure how relevant each document is to an intent, weighted by the importance of each intent;
Table 5: Effectiveness results based on the I-rec, D-nDCG and D#-nDCG measures for the subtopic mining run RMIT-S-J-1.

<table>
<thead>
<tr>
<th>Run</th>
<th>I-rec@10</th>
<th>D-nDCG@10</th>
<th>D#-nDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMIT-D-J-4</td>
<td>0.6489</td>
<td>0.3301</td>
<td>0.4895</td>
</tr>
<tr>
<td>RMIT-D-J-3</td>
<td>0.6723</td>
<td>0.3664</td>
<td>0.5193</td>
</tr>
<tr>
<td>RMIT-D-J-1</td>
<td>0.6356</td>
<td>0.3540</td>
<td>0.4948</td>
</tr>
<tr>
<td>RMIT-D-J-2</td>
<td>0.6306†</td>
<td>0.3283</td>
<td>0.4795</td>
</tr>
<tr>
<td>RMIT-D-J-5</td>
<td>0.5639‡</td>
<td>0.2989</td>
<td>0.4314‡</td>
</tr>
</tbody>
</table>

Table 6: Effectiveness results based on the I-rec, D-nDCG and D#-nDCG measures at cutoff 10. Run RMIT-D-J-4, which uses the Okapi BM25 similarity measure with the original queries and no diversification of the ranked results list is treated as a baseline; † and ‡ indicate statistical significance relative to the baseline at the 0.05 and 0.001 levels, respectively, based on a paired t-test.

4.2 Subtopic Mining

For the subtopic mining task, we submitted a single run, RMIT-S-J-1. The effectiveness results are shown in Table 5. Our run for the subtopic mining task did not identify many subtopic strings. Consequently, our intent recall was low. Figure 2 shows the number of identified subtopics strings and intents in our runs, together with the number of officially judged intents in the task.

4.3 Document Ranking

Five runs were submitted by our team for the intent document ranking task, as detailed in Table 3. The run RMIT-D-J-4 used the Okapi BM25 similarity function and the original topic queries, segmented using MeCab (Vanilla 1). We therefore consider this run as a baseline, against which the effectiveness of our other runs is compared.

Tables 6 and 7 show the results of our runs for the three evaluation metrics with cutoffs at rank positions 10 and 30, respectively. For both cutoff levels, the best performance was achieved by run RMIT-D-J-3 which used the Diversity 1 queries with TF weighting, and a round-robin re-ranking approach for diversification of the results list, leading to an absolute improvement in D#-nDCG of 0.0298 and 0.0004 for cutoff levels 10 and 30, respectively. However, these improvements are not statistically significant relative to the RMIT-D-J-4 baseline using Okapi BM25 and no diversification.

Adding round-robin diversification to the BM25 similarity measure (run RMIT-D-J-1) led to a fractional improvement over the non-diversified baseline at cutoff level 10 (an absolute increase in D#-nDCG of 0.0053), and a fractional decrease in performance relative to the baseline at cutoff level 30 (an absolute decrease in D#-nDCG of 0.013); the impact of this diversification approach on the Okapi BM25 measure was not statistically significant in either case.

The additional processing of query terms by dropping non-nouns (Vanilla 2 queries in run RMIT-D-J-2) led to a statistically non-significant decrease in Okapi BM25
Table 7: Effectiveness results based on the I-rec, D-nDCG and D#-nDCG measures at cutoff 30. Run RMIT-D-J-4, which uses the Okapi BM25 similarity measure with the original queries and no diversification of the ranked results list, is treated as a baseline; † and ‡ indicate statistical significance relative to the baseline at the 0.05 and 0.001 levels, respectively, based on a paired t-test.

<table>
<thead>
<tr>
<th>Run</th>
<th>I-rec@30</th>
<th>D-nDCG@30</th>
<th>D#-nDCG@30</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMIT-D-J-4</td>
<td>0.8012</td>
<td>0.3617</td>
<td>0.5814</td>
</tr>
<tr>
<td>RMIT-D-J-3</td>
<td>0.7836</td>
<td>0.3800</td>
<td>0.5818</td>
</tr>
<tr>
<td>RMIT-D-J-2</td>
<td>0.7977</td>
<td>0.3575</td>
<td>0.5776</td>
</tr>
<tr>
<td>RMIT-D-J-1</td>
<td>0.7752</td>
<td>0.3617</td>
<td>0.5684</td>
</tr>
<tr>
<td>RMIT-D-J-5</td>
<td>0.6759†</td>
<td>0.3118‡</td>
<td>0.4938‡</td>
</tr>
</tbody>
</table>

Finally, combining the Vanilla 2 queries with a TF weighting scheme (run RMIT-D-J-5) led to the lowest performance for both cutoff levels, with a statistically significant decrease relative to the baseline run.

5. CONCLUSIONS

In this report, we have presented results for our new experimental ranked self-index NeWT, using the ClueWeb09-JA collection. Our experiments investigated the impact of diversifying topics using the Japanese Wikipedia, and using a round-robin approach for re-ranking search result lists for individual topic aspects. These approaches did not have a significant impact on search performance as measured by D#-nDCG, relative to an already strong BM25 baseline. We suspect that this is due to the relatively small number of intents that were identified, compared to the total number of officially identified intents (see Figure 2).

While overall effectiveness for document ranking in our first attempt with the INTENT topics is suboptimal, we have shown that ranked self-indexes are a viable alternative to classical inverted indexing approaches for Gigabyte scale Japanese document collections. Furthermore, our method does not require any domain knowledge about the underlying text being indexed, allowing all domain and language decisions to be deferred until query time. In future work, we will investigate new approaches to improving system effectiveness using query expansion and alternative ranking algorithms within our self-indexing framework.

6. ACKNOWLEDGMENTS

The second author was supported by the Australian Research Council.

7. REFERENCES


