In this article, we assess the effectiveness of Contextual Document Clustering (CDC) as a means of indexing within a dynamic and rapidly changing environment. We simulate a dynamic environment, by splitting two chronologically ordered datasets into time-ordered segments and assessing how the technique performs under two different scenarios. The first is when new documents are added incrementally without reclustering [incremental CDC (iCDC)], and the second is when reclustering is performed [nonincremental CDC (nCDC)]. The datasets are very large, are independent of each other, and belong to two very different domains. We show that CDC itself is effective at clustering very large document corpora, and that, significantly, it lends itself to a very simple, efficient incremental document addition process that is seen to be very stable over time despite the size of the corpus growing considerably. It was seen to be effective at incrementally clustering new documents even when the corpus grew to six times its original size. This is in contrast to what other researchers have found when applying similar simple incremental approaches to document clustering. The stability of iCDC is accounted for by the unique manner in which CDC discovers cluster themes.

Introduction

Document clustering in Information Retrieval provides numerous benefits: It can improve the quality of query-based retrieval (Liu & Croft, 2004; Tombros, Villa, & Van Rijsbergen, 2002), it provides a topical structuring organization of the corpus (Dobrynin, Patterson, Galushka, & Rooney, 2005), and it facilitates the browsing of the corpus (Cutting, Karger, Pedersen, & Tukey, 1992). However, research into document clustering methods for indexing a corpus and retrieval is often only considered under the assumption that the corpus is static. This is an unlikely scenario in real-world IR search systems, where a document corpus is dynamic in the sense that the documents can be added to or deleted from the corpus. Such processes can significantly modify the content of the corpus and therefore the optimal cluster structure. This has implications in terms of the effectiveness of document clustering methods, in that the resultant clusters cannot be easily maintained and tend to degrade in quality with document additions/deletions. Of course, the problem can be addressed by reclustering the whole collection. But for large collections, this is computationally expensive and infeasible for some clustering methods such as agglomerative clustering. In general, the maintenance strategy cannot be isolated from the nature of clustering. A simple strategy such as assigning a new document to the cluster it is most similar to depends on what determines “similarity,” which in general will be integral to the clustering mechanism (e.g., cosine similarity in a vector-based k-means (Steinbach, Karypis, & Kumar, 2000) or KL-divergence in language models (Ponte & Croft, 1998). A simple maintenance strategy such as this has been found to degrade the quality of the clustering structure even with relatively small increases in the number of new documents (i.e., from 25% of corpus size) (Can, Fox, Snavely, & France, 1995).

An alternative approach to the problem of cluster maintenance is to adopt a clustering algorithm that is by its nature incremental. Such an algorithm should allow the addition of new documents without the requirement for reclustering while maintaining the cohesiveness of existing clusters and allowing for the creation of new clusters or merging of existing clusters if required. Other issues of importance include whether the performance of the algorithm is dependent on the insertion order of documents and whether a document can be reassigned to another cluster. We distinguish between incremental clustering based on the bag of words model and
phrase-based models. Bag of words based techniques can be
further subdivided based on whether they are partitional or
hierarchical in nature. Partitional methods include single-
pass clustering, nearest neighbor clustering (M.A. Wong &
Lane, 1983), cover coefficient concept clustering Can et al.
(1995), and the artificial neural network technique of adap-
tive resonance theory (ART; Massey, 2003). The most simple
strategy for incremental clustering is single-pass clustering,
where documents are assigned to an existing cluster if they
exceed a similarity threshold; otherwise, they form a cluster
of their own. In nearest neighbor clustering, a document is
assigned to the cluster, which has a majority of its nearest
neighbors. The cover coefficient concept proposed by Can
(1993) adopts a strategy where a cover coefficient matrix is
constructed, and each entry is a measure of how much the
document is “covered” by itself in the case of diagonal ele-
ments and by another document in the case of off-diagonal
elements. This matrix allows a single-pass clustering algo-

rithm to be developed by calculating the seed power of
documents based on the cover coefficient matrix. Clusters
are formed, seeded by documents with the highest seed
power, and documents are assigned to clusters based on the
seed which covers them maximally. Clusters can be easily
updated by identifying whether a new document or old non-
seed document constitutes a new seed document and
whether any of the previous seed documents should be
considered as no longer a seed. The latter can lead to clusters
being deconstructed and their contents being reassigned to
new/existing clusters formed by the seed documents. ART
neural networks have the ability to provide online and incre-
mental clustering of dynamic datasets and have recently
been considered for text clustering; however, their ability
to provide incremental document clustering has not been
extensively evaluated. Hierarchical methods include the
standard hierarchical agglomerative clustering (HAC) tech-
nique, and DC-tree clustering. HAC clustering, although it
provides a means of providing an incremental clustering
based on single link, is not considered very effective (Can,
document clustering approach called DC-tree clustering,
which has parallels to B+-tree indexing except that the DC-
tree is not height balanced, and each leaf node represents a
cluster and not a document. They use three distance-based
measures for similarity, which they considered more appro-
piate than the more commonly used vector space model.

A number of Web document clustering methods have
been proposed based on phrasal-based incremental clus-
tering. Zamir and Etzioni (1998) described an algorithm refe-
t to as suffix-tree clustering, which builds a tree of phrase
suffixes between multiple documents. This results in an in-
verted index of base clusters of documents containing the
same phrase. An agglomerative single-link clustering
algorithm is used to combine base clusters based on a binary
similarity measure. Hammouda and Kamel (2004) also
adopted a phrasal approach to clustering. They created a doc-
ument index model, the Document Index graph, which facil-

itates incremental construction of a phrase-based document
index. They build cluster similarity histograms, which repre-
sent the pairwise document similarity distribution within the
cluster. They try to maintain the maximization of the number
of similarities in the similarity intervals for each cluster, for
the addition of a new document.

In this article, we assess the stability of Contextual Doc-
ument Clustering (CDC; Dobrynin, Patterson, & Rooney,
2004) in the context of dynamic and rapidly changing
domains where new documents are continually being
added. We examine how robust the technique is using a very
simplistic, low-cost, update policy where new documents
are added to the cluster that they are most similar to, with a
view to determining if any loss in competency occurs, as
was reported by other researchers using this simplistic
approach to maintenance. We also compare this simple
maintenance approach to reclustering the entire corpus on a
regular basis.

Methodology

The key concept in CDC is the identification of implicit
themes within a corpus, which are of such specificity for
grouping a relatively small number of documents into a clus-
ter, all of which are related to the theme. A theme is a proba-
bility distribution of co-occurring terms given a chosen
theme term and provides the means of forming clusters. Of
interest are theme terms which form specific themes occur-
ing, for example, in a document collection; the theme term
“processor” may be specific in scope, but the term “com-
puter” is unlikely to be as it can occur in many different
contexts. The process by which themes are discovered is
described in detail in Dobrynin et al. (2003) and Dobrynin
et al. (2005). Of interest in this study is determining how
stable these themes are in environments that are updated
regularly. Here, the concept of stability is based on assum-
ing that a theme once identified has longevity, in the sense
that as new documents are added to the collection, estab-
lished themes are still relevant and still allow for coherent
clusters. The opposite contention is that themes are unstable,
and it takes a relatively small volume of documents to be
added to clusters to degrade the quality of clusters, and that
it would be better to rediscover themes from first principles.
As such, CDC would not be considered as a suitable ap-

proach for dynamic environments. To determine whether
this is so, we benchmark this process of incremental clus-
tering via static themes to the process of nonincremental
reclustering, where after each batch addition of documents
we rediscover themes and recluster. Incremental clustering
via static themes discovers themes only once with the initial
starting batch of documents, then it adopts an efficient
single-pass clustering method for dynamic additions, where
a new document is assigned to a cluster based on which clus-
ter theme that it is most similar to. The nonincremental
approach rediscover themes and reclusters after every batch
addition. Internally within a cluster, documents are orga-
nized in the form of a minimum spanning tree (MST), where
each edge length between adjacent documents is based on
According to Heap’s Law (Baeza-Yates & Ribeiro-Neto, 2000), which contain the term $z$, where $z$ represents the term $z$ in a randomly selected document within which the term $z$ co-occurs. We can approximate this distribution as:

$$p(y | z) = \frac{\sum_{x \in X(z)} tf(x, y)}{\sum_{x \in X(z), t \in Y} tf(x, t)}$$

where $tf(x, y)$ is the term frequency of the term $y$ in document $x$ and $X(z)$ is the set of all documents from the corpus which contain the term $z$.

It is obvious that in most cases, the context of the term $z$ is too general in scope to present useful information about the corpus. So we are interested only in identifying terms $z$ which identify narrow contexts. It is these narrow contexts we refer to as themes. Themes are identified by a consideration of the entropy of the probability distribution $p(Y | z)$:

$$H(Y | z) = - \sum_y p(y | z) \log(p(y | z))$$

and the document frequency of the context term $df(z)$.

Let $Y(z)$ denote the set of all the different terms from documents in $X(z)$. When there is a uniform distribution of terms from $Y(z)$, the entropy $H(Y | z)$ is equal to $\log |Y(z)|$. According to Heap’s Law (Baeza-Yates & Ribeiro-Neto, 2001), $\log |Y(z)| = O(\log (|X(z)|))$ (i.e., the entropy of a term $z$ is related to the log of the size of the set of documents containing the term $X(z)$. As the document frequency of the term $z$, $df(z)$ equals $|X(z)|$, the entropy of the context is related to the logarithm of the document frequency of its context term. Therefore, high document frequency terms have higher entropy values than do low document frequency terms. This means that it is insufficient to choose themes based on a consideration of their entropy alone. To take into account the dependency between the document frequency $df(z)$ of the term $z$ and the entropy of its context, we divide context terms into disjoint subsets based on their document frequency:

$$Y' = \bigcup_i Y_i$$

$$Y_i = \{z : z \in Y, df_i \leq df(z) < df_{i+1}\}$$

$i = 1, \ldots, r$

Here, the threshold $df_i$ satisfies the condition $df_{i+1} = \alpha df_i$, where $\alpha > 1$ is a constant. The choice of document frequency thresholds is based on the following principles. As in other IR applications, we consider context terms that have very low or high document frequency as not being informative; hence, such terms are not considered as being potential thematic terms (Sebastiani, 2002). Removing such terms has a long history in Information Retrieval (Luhn, 1958), and determining whether terms were based as too “infrequent” or too “frequent” is arbitrary and often carried out based on trial and error (Van Rijsbergen, 1979). In our process for removing frequent/infrequent terms, we ranked each term by its document frequency and excluded from consideration the bottom very infrequently occurring terms and the top very frequent terms. The remaining terms we divided into $r$ document frequency intervals, where the interval bounds $df_i$ were set based on setting $\alpha = 2$.

We assume in total that there are at most $N$ narrow theme terms. For every $i = 1, \ldots, r$, a set $Z_i \subset Y_i$ is selected such that:

$$|Z_i| = \frac{N |Y_i|}{\sum_{j=1, \ldots, r} |Y_j|}$$

and

$$z_1 \in Z_1, z_2 \in Y_i - Z_4 \rightarrow H(Y | z_1) \approx H(Y | z_2).$$

As such, we select from each subset $Y_i$, a proportionate number (proportional to the size of subset $|Y_i|$ relative to the size of the union of all subsets $Y'$) of theme terms that fall into this interval and which have the highest entropy.

Next, we perform a theme-merging process to ensure that themes that are very similar do not compete for the same documents during clustering. Given a set of narrow theme terms $Z = \bigcup_i Z_i$, we merge some of the themes into one theme if the distance between them is less than a fixed threshold. Given two theme terms $z_1$ and $z_2$, the distance between the two themes $p(Y | z_1)$ and $p(Y | z_2)$ is based on the JS divergence between the two themes:

$$JS_{0.5, 0.5}[p(Y | z_1), p(Y | z_2)] = \frac{H[p] - 0.5H[p_1] - 0.5H[p_2]}{H[p] - H[p_1] - H[p_2]}$$

Initially, all narrow themes are ordered by theme ID. Starting with the first theme (i.e., the seed theme), all remaining themes are sequentially considered for merging based on their distance from the seed. If a theme’s distance is sufficiently small (i.e., less than the set threshold), then it is merged with the seed. Once this first pass is finished, all themes now grouped with the seed theme are removed from $Z$ and are collectively described by one new theme, which is the average for this new merged group. The next theme ID, still within $Z$, now becomes the next seed theme candidate to form merged themes, and all remaining themes in $Z$ are considered sequentially based on their distance to it, as before. This process continues until all themes in $Z$ are considered as seeds for merged themes. Clustering then follows a single-pass process of assigning a document to the “closest” cluster with theme term $z$ by comparing the distance, as measured by the JS divergence.
between itself and each cluster’s theme distribution (i.e., this is a hard clustering approach). The “closest” cluster to a document \(x\) is the one with the minimum JS divergence between the document’s probability distribution and the cluster’s theme.

\[
z = \arg \min_{1 \leq z \leq T} JS_{[0.5,0.5]}[p(Y|x), p(Y|z)]
\]

**Experimental Evaluation**

The process of comparing iCDC to nCDC is based on two datasets which contain chronologically ordered data. The first dataset is the Reuters Corpus Volume 1 (RCV1; Rose, Stevenson, & Whitehead, 2002), and the second dataset is the OHSUMED collection (Hersh, Buckley, Leone, & Hickman, 1994). Both collections contain topical categories assigned to individual document collections. This independent information is not used in the CDC process, but does allow us to evaluate the quality of clustering. We measure cluster quality using two different measures: Jaccard coefficient (JC) and accuracy of document classification.

**JC**

The JC is measured by calculating the topic overlap between pairs of adjacent documents. A document \(x_1\) is adjacent to another \(x_2\) if they share an edge in the minimum spanning tree within a cluster (Patterson, Rooney, Galushka, & Dobrynin, 2005). The minimum spanning tree distance measure is based on the JS divergence between documents. If \(T(x_1)\) is the topic set of \(x_1\) and \(T(x_2)\) is the topic set of \(x_2\), then the JC can be measured as:

\[
JC = \frac{|T(x_1) \cap T(x_2)|}{|T(x_1) \cup T(x_2)|}
\]

The topic set for a document is the set of categories assigned to it.

**Document Classification**

We consider the following classification framework to measure the quality of clusters formed. Cluster Classification (CC) assesses the categories of all documents in a cluster and assigns documents the categories whose popularity within this cluster is above a defined threshold. Let \(T(C; \beta)\) be the set of all categories assigned by experts to documents within a Cluster \(C\), whose popularity (where popularity is defined as the percentage of documents having this category) exceeds Threshold \(\beta\). Then Classifier CC classifies every Document \(d\) from the cluster as category set \(T(C; \beta)\). We measure the strength of this classification approach using standard measures of micro-averaged precision \(p\) and recall \(r\), and the harmonic mean, \(F\) measure:

\[
F - \text{measure} = \frac{2pr}{p + r}
\]
We consider terms that occur in Interval 0 and terms that occur in Interval 6 as occurring either too frequently or too infrequently, respectively, to act properly as theme terms. As such, we define five document-frequency intervals whose boundaries vary depending on the size of the collection. Taking into account that the number of documents published in any given month is approximately equal to 1/12 of the total collection size, the document-frequency intervals were set based on \((\text{df}_i \times k/12, \text{df}_i/12)\), where \(k\) represents the number of month’s worth of documents used to discover themes thus far and \(i\) is one of the five given intervals. Table 2 shows the interval settings for each technique.

### JC

We determined the JC for each pair of adjacent documents for each category set and determined the frequency of documents falling into 1 of 10 equal-sized intervals, where the lowest interval is (0.0,0.1) and the highest interval is from (0.9,1.0). Note a value of 1.0 complete disagreement.

Figure 1 shows the initial distribution of the proportion of pairs for JC intervals, after the Month 1 for both iCDC and nCDC (Note that they are identical for Month 1.). Clearly, for Topics and Regions, the majority of cases fall into the highest interval. This is not the case for Industries, which has been shown to be a difficult classification set (Lewis et al., 2004). From these graphs, it can be seen that 50% of document pairs for Topics and 60% for Regions fall into the highest interval. Additionally, it can be seen that 72% of document pairs have a very good result considering that there are 103 different Topic categories, 354 different region categories, and a document can have more than one category (On average, each document has 3.241 categories in Topics and 1.315 categories in Regions.)

Having examined the initial state (Month 1), we now investigate the impact on JC of the two variations of CDC over time. Figure 2 shows the fraction of document pairs falling into the highest JC interval (0.9,1.0) for the first 6 months, for the three category sets, for nCDC and iCDC (The exact values are shown in Table 7 in the Appendix.)

The graphs show that the results for iCDC are remarkably stable; there is no noticeable deterioration in JC values for any of the three category sets, as new documents are accumulated over time. In fact, an improvement can be observed over time. nCDC also shows improvement in JC overlap. These improvements over time are only marginally better than those of iCDC. For example, with Topics at Month 6, the difference between the two clustering approaches is only 0.01, with Regions it is 0.05, and with Industries it is 0.03.

Clearly, nCDC will result in more clusters being formed over time in comparison to iCDC due to the fact that more documents and terms are considered during theme identification. This can be seen in Figure 3, which shows that nCDC forms around 500 more clusters than does iCDC by Month 6.

---

### TABLE 1. Document frequency intervals for the RCV1-v2* collection.

<table>
<thead>
<tr>
<th>Interval ((i))</th>
<th>(\text{df}<em>i, \text{df}</em>{i+1})</th>
<th>No. of terms</th>
<th>%/total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(1, 60)</td>
<td>476,287</td>
<td>0.938</td>
</tr>
<tr>
<td>1</td>
<td>(60, 120)</td>
<td>10,410</td>
<td>0.020</td>
</tr>
<tr>
<td>2</td>
<td>(120, 240)</td>
<td>7,231</td>
<td>0.014</td>
</tr>
<tr>
<td>3</td>
<td>(240, 480)</td>
<td>4,966</td>
<td>0.010</td>
</tr>
<tr>
<td>4</td>
<td>(480, 960)</td>
<td>3,143</td>
<td>0.006</td>
</tr>
<tr>
<td>5</td>
<td>(960, 1920)</td>
<td>1,987</td>
<td>0.004</td>
</tr>
<tr>
<td>6</td>
<td>(\geq 1920)</td>
<td>3,859</td>
<td>0.008</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>507,883</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 2. Document frequency intervals for each technique.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Interval settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>iCDC</td>
<td>([\text{df}<em>i/12, \text{df}</em>{i+1}/12])</td>
</tr>
<tr>
<td>nCDC</td>
<td>([\text{df}<em>i \times k/12, \text{df}</em>{i+1} \times k/12])</td>
</tr>
</tbody>
</table>

---

1Note the number of clusters formed using iCDC also increases slightly. This is not due to the discovery of new themes, as with nCDC, but the utilization of previously empty themes by assigning new documents to them.
Clearly, nCDC will form clusters of smaller average size over time than will iCDC as there are more clusters and the same number of documents. This improves the average homogeneity of clusters leading to slightly improved JC overlaps. Further evidence for this is seen from examining the measure used for assessing similarity, the JS divergence. It can be seen from Figure 4 that the average distance between adjacent documents in the MST decreases more with nCDC than it does with iCDC, indicating that with nCDC the clusters are actually becoming more homogeneous in terms of their similarity over time. This ties in with the earlier observation that JC values are slightly higher for nCDC.

Tables 3 to 5 show the proportion of document pairs falling into the lowest and highest intervals for both techniques after 6 months of data had been clustered, for each category set. Theoretically, an optimal clustering algorithm will have no document pairs falling into the lowest JC interval and all document pairs falling into the highest JC interval. The results emphasize the quality of CDC, especially for Topics and Regions and how closely matched the two techniques are. Although nCDC consistently outperforms iCDC in both intervals (0.0, 0.1) and (0.9, 1.0), it is always only by a small amount.

### Table 3. Fraction of documents falling into the lowest and highest intervals for the Topics set.

<table>
<thead>
<tr>
<th>Technique</th>
<th>JC interval [0.0,0.1)</th>
<th>JC interval [0.9,1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCDC</td>
<td>0.05</td>
<td>0.54</td>
</tr>
<tr>
<td>iCDC</td>
<td>0.07</td>
<td>0.53</td>
</tr>
</tbody>
</table>

### Table 4. Fraction of documents falling into the lowest and highest intervals for the Regions set.

<table>
<thead>
<tr>
<th>Technique</th>
<th>JC interval [0.0,0.1)</th>
<th>JC interval [0.9,1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCDC</td>
<td>0.13</td>
<td>0.70</td>
</tr>
<tr>
<td>iCDC</td>
<td>0.18</td>
<td>0.66</td>
</tr>
</tbody>
</table>

### Table 5. Fraction of documents falling into the lowest and highest intervals for the Industries set.

<table>
<thead>
<tr>
<th>Technique</th>
<th>JC interval [0.0,0.1)</th>
<th>JC interval [0.9,1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCDC</td>
<td>0.46</td>
<td>0.34</td>
</tr>
<tr>
<td>iCDC</td>
<td>0.51</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Classification

Figure 5 shows the results, in terms of the $F$ measure, of the CC classifier for the three category sets (actual $F$-measure values also are shown in Table 8 in the Appendix.) Regarding Topics and Regions, nCDC rose in value with each additional month, showing an overall improvement in $F$ measure of 0.04 for Topics and 0.1 for Regions. For the same category sets, iCDC tended to remain constant with Topics and to drop a little with Regions (by 0.02), which would suggest that the iCDC clusters do not degrade in their level of cohesiveness (even if they do not necessarily improve). The results for the Industries set is somewhat anomalous. nCDC shows only a small improvement (0.02) whereas for iCDC, the $F$ measure does decrease (0.05 in total). There is evidence from previous studies that it is difficult to obtain consistent results for this category set (Lewis et al., 2004). Overall, it can be concluded that despite forming themes based on only 62,582 documents and incrementally adding 322,113 new documents over time, CDC does not lose significant quality. This is in contrast to what researchers have reported with other clustering approaches Can et al. (1995), where a degradation in quality is observed with this simple maintenance policy even with the addition of only a small percentage of new documents. Considering the obvious efficiency advantages of iCDC to nCDC, it is therefore an attractive solution for clustering large corpora that are dynamic in nature. Additionally, these results indicate the stability of themes over time.

OHSUMED Collection

We repeated our evaluation using the TREC-9 filtering task derivative of the OHSUMED collection, first presented to the research community by Hersh et al. (1994). The OHSUMED collection consists of over 300,000 MEDLINE documents collated from 270 medical journals published in the Years 1987 to 1991. A total of 233,445 documents, those which contained both a title and an abstract, were used in the evaluation. The number of documents for each year is approximately similar. Each document has a title, a set of expert assigned MeSH descriptors, and an abstract. Only the abstracts and titles were used in the clustering process. The MeSH descriptors allowed us to independently assess the quality of clustering. We set the document-frequency intervals so that the same proportion of documents fell into each interval as in RCV1-v2*, as shown in Table 6. Rather than divide the data into segments based on months, in this instance, we divided the data based on years where Year 1 corresponds to 1987, and so on.

As with RCV-v1, we considered only the document frequency intervals for $I = 1, \ldots, 5$. The document-frequency intervals were set for nCDC for each additional $k$th year addition as $(df_i * k/5, df_i + 1 * k/5)$.

iCDC was based on discovering themes from the first year of data (1987), approximately 46,000 documents, and using these themes to cluster documents for the remaining 4 years in incremental batch fashion. Conversely, nCDC discovers themes based on the documents currently clustered plus the following year’s documents. It then reclusters all documents based on the new theme set. For both iCDC and nCDC, the value of $N$ was set to 2000.

JC

Figure 6 shows the proportion of documents falling into different JC coefficient intervals for both nCDC and iCDC after 1 year of data.

Clearly, a far larger proportion of adjacent document pairs have a low or low-medium JC coefficient in this collection of only a small percentage of new documents. Considering the obvious efficiency advantages of iCDC to nCDC, it is therefore an attractive solution for clustering large corpora that are dynamic in nature. Additionally, these results indicate the stability of themes over time.
than those in the Reuters collection. This is a consequence of the fact that as opposed to the case for RCV1 where documents were assigned on average 3.24 categories of 103 for Topics, for example, here documents have been assigned many more MeSH terms on average (12.32) of a total number of 14,334 terms. Naturally, as a consequence, topic similarity is reduced.

Note that the focus in these experiments is not how well the clustering approach performs in terms of classification (It has been previously demonstrated to be a competent approach for both retrieval and classification; Dobrynin et al., 2005). More important is the comparison between the relative performances of nCDC and iCDC over time in terms of the proportion of documents falling into the different JC intervals.

Figure 7 shows that with increasing years, the proportion of document pairs falling in the lowest interval of topic overlap decreases with both approaches, although the improvement is very slightly greater with nCDC than with iCDC. By Year 5, the difference between iCDC and nCDC is 0.01.

Figure 8 shows the resultant JC interval distributions for both iCDC and nCDC, overlaid with the initial CDC distribution. This again highlights the improvement observed across all JC intervals. It can be seen that there is a shift away from the lower intervals toward the upper intervals and that there is no noticeable difference between iCDC and nCDC.

With the OHSUMED collection, the comparison in the number of clusters formed between iCDC and nCDC is more consistent (in comparison to Reuters), as on initial inspection the number of clusters formed by each clustering technique for each year is roughly the same (see Figure 9). This would initially imply that no new themes are being discovered over time by nCDC with OHSUMED. This in fact is not the case. New themes may be discovered and replace existing themes; however, the total overall themes discovered is bounded by the upper limit on the number of themes to use.

Classification

Figure 10 shows the results of the CC classifier for both nCDC and iCDC. F values also are shown in Table 9 in the Appendix. Not unexpectedly, nCDC performs better than
does iCDC; however, the difference in $F$ measure between the techniques is not large and is always less than 0.01, even in Year 5. Both techniques showed a very small, but detectable, decrease in $F$ measure over time (0.1 for nCDC and 0.2 for iCDC). This is in contrast to the Reuters results, where a slight increase or stability was observed in $F$ measure for Topics and for Regions. The reason for this is possibly the limit to the number of themes discovered (2,000), and consequently, the number of clusters created is too low, as the collection is incrementally increased in size.

**Topic Change**

Of interest is the degree to which topic changes occur over the time frame assessed for the RCV1-v2* and the OHSUMED collection to clarify whether the results for iCDC are consistent with this issue. This is difficult to assess for OHSUMED, as the number of possible topics is large (discussed earlier) and many may not be independent. In the case of RCV1-v2*, this was possible allowing that we only considered category sets that are independent and general enough to occur in a sufficient number of documents. For the Topic category set, we chose categories from the second level of the hierarchy. For Industries, we chose the fifth level of categories. All categories were considered in the case of “Regions.” We considered the first 34 weeks of data using 1 week as a unit of time, which seemed appropriate for news stories.

We wished to consider two issues: (a) how much the probability distribution of topics itself changes over the time period and (b) how different the probability distribution of a given topic is to a uniform distribution over the time period.

$JS$ divergence provides a natural means of measuring the variation in topic distributions week by week. Figures 11 to 13 show the $JS$ divergence between Week $i$ and Week $i+1$ for the three category sets. In the case of the Topics category set, a noticeable jump in the $JS$ divergence occurs in the case of Week(2,3), Week(3,4), Week(19,20), Week(20,21), Week(30,31), and Week(31,32), indicating for 3 weeks the occurrence of some distinct news event. In the case of Industries, there were $JS$ divergence spikes for the same week sets, but the values of $JS$ for all other weeks were generally higher, indicating a greater level of topical change. The Regions category set follows the same behavior as Topics.

The second issue of determining how different the probability distribution of a given topic is to uniform distribution over the time period was assessed by determining the normalized frequency of occurrence for each category (or probability of occurrence) and determining the standard deviation in normalized frequency of occurrence over the 34 weeks.

Figures 14 to 16 show the variation in standard deviation for a representative number of categories in each category.
As the probability of occurrence for each week of a given category sums to 1 over the 34 weeks, the mean probability is always 1/34 or 0.029, regardless of the category. The ratio of the standard deviation to the mean gives the coefficient of variation (CV). If the CV is less than 1, the distribution is considered to have low variance. It can be seen that only in the case of Industries or Regions are there categories which have a CV of 1 or close to it; however, these categories are not in the majority.

In summary, the JS divergence and standard deviation measures show a small degree of topical variation with the Topics and Regions sets, and a much pronounced variation within the Industries set, so the JC and classification results for nCDC and iCDC were recorded in the presence of a certain degree of topical change in the case of RCV1.

Conclusions

The general conclusion from this study is that for both datasets, the nonincremental technique version of CDC performed marginally, yet consistently, better than the incremental version; however, note that there was never a noteworthy performance difference. Important, the incremental version did not show a drop in performance, as is often seen with other clustering approaches employing iCDC’s simplistic maintenance strategy. This indicates that the quality of the clusters formed by incremental addition was robust, relative to the nonincrementally formed clusters. This is shown both for the RCV1-v2* collection, where the additional documents increased the collection size to almost six times that of the original corpus, and for the OHSUMED collection, where the additional documents increased the collection in size by approximately five times its original amount. These are significant increases in corpus sizes for the incremental approach to deal with, without rediscovering new themes and reclustering the documents (as occurred with the nonincremental approach). This is a significant result, as clustering approaches are notorious for degrading in quality as new data is added to the dataset, in the absence of reclustering.

As such, our view is that CDC is able to identify stable themes from a small proportion of a dataset (1 month in the case of 6 months of data from RCV1-v2* and 1 year in the case of 5 years of data from OHSUMED), which facilitates a simple incremental clustering mechanism, which performs to an acceptably high level and is a very close approximation to a nonincremental version of the same algorithm. The advantage of the incremental approach is that documents can be added in real time, and there is an obvious efficiency saving by virtue of circumnavigating the need to constantly recluster the corpus. The reason iCDC is so effective is due to the unique manner in which themes for the clusters are discovered. Unlike other clustering approaches, themes represent important concepts which naturally occur within corpora. In general, these themes are very stable and highly relevant over time, even within a rapidly changing and dynamic environment such as news stories. Therefore, as long as a sufficient number of documents are used initially to discover the themes, the number of new themes introduced by the addition of new documents over time is limited, and as such, these have a relatively small impact on the overall quality of the approach. Inevitably of course, given enough time, the quality of the
system will benefit from rediscovering a new set of themes. We have demonstrated in this work that the corpus can grow to more than six times its original size (in the case of RCV1) without a significant loss in quality. To our knowledge, this is the largest study of its kind (in terms of the number of documents used: 384,695 documents from RCV1) in assessing the incremental nature of a clustering algorithm. This demonstrates the scalability of the CDC approach and its potential application to real-world corpora.

References


Appendix

TABLE 7. JC values in interval [0.9,1.0] for RCV1 for the first 6 months.

<table>
<thead>
<tr>
<th>Month</th>
<th>Topics icDC</th>
<th>nCDC</th>
<th>Regions icDC</th>
<th>nCDC</th>
<th>Industries icDC</th>
<th>nCDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.504</td>
<td>0.504</td>
<td>0.621</td>
<td>0.621</td>
<td>0.277</td>
<td>0.277</td>
</tr>
<tr>
<td>2</td>
<td>0.512</td>
<td>0.518</td>
<td>0.648</td>
<td>0.629</td>
<td>0.286</td>
<td>0.281</td>
</tr>
<tr>
<td>3</td>
<td>0.532</td>
<td>0.543</td>
<td>0.681</td>
<td>0.649</td>
<td>0.306</td>
<td>0.282</td>
</tr>
<tr>
<td>4</td>
<td>0.531</td>
<td>0.543</td>
<td>0.686</td>
<td>0.652</td>
<td>0.323</td>
<td>0.295</td>
</tr>
<tr>
<td>5</td>
<td>0.533</td>
<td>0.547</td>
<td>0.692</td>
<td>0.653</td>
<td>0.333</td>
<td>0.304</td>
</tr>
<tr>
<td>6</td>
<td>0.528</td>
<td>0.544</td>
<td>0.702</td>
<td>0.658</td>
<td>0.339</td>
<td>0.311</td>
</tr>
</tbody>
</table>

TABLE 8. F-measure values for RCV1 for the first 6 months.

<table>
<thead>
<tr>
<th>Month</th>
<th>Topics icDC</th>
<th>nCDC</th>
<th>Regions icDC</th>
<th>nCDC</th>
<th>Industries icDC</th>
<th>nCDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.637</td>
<td>0.637</td>
<td>0.521</td>
<td>0.521</td>
<td>0.362</td>
<td>0.362</td>
</tr>
<tr>
<td>2</td>
<td>0.634</td>
<td>0.658</td>
<td>0.509</td>
<td>0.556</td>
<td>0.336</td>
<td>0.346</td>
</tr>
<tr>
<td>3</td>
<td>0.641</td>
<td>0.668</td>
<td>0.514</td>
<td>0.574</td>
<td>0.315</td>
<td>0.354</td>
</tr>
<tr>
<td>4</td>
<td>0.636</td>
<td>0.676</td>
<td>0.505</td>
<td>0.593</td>
<td>0.309</td>
<td>0.343</td>
</tr>
<tr>
<td>5</td>
<td>0.634</td>
<td>0.683</td>
<td>0.502</td>
<td>0.62</td>
<td>0.312</td>
<td>0.38</td>
</tr>
<tr>
<td>6</td>
<td>0.635</td>
<td>0.683</td>
<td>0.501</td>
<td>0.62</td>
<td>0.306</td>
<td>0.384</td>
</tr>
</tbody>
</table>

TABLE 9. F-measure values for OHSUMED for the first 6 months.

<table>
<thead>
<tr>
<th>Year</th>
<th>icDC</th>
<th>nCDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.444</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.431</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.426</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.421</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.417</td>
<td></td>
</tr>
</tbody>
</table>