Analysis of Source Identified Text Corpora: Exploring the Statistics of the Reused Text and Authorship

Akiko Aizawa
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, 101-8430, Japan
akiko@nii.ac.jp

Abstract

This paper aims at providing a view of text recycled, within a short time, by the authors themselves. We first present a simple and general method for extracting reused term sequences, and then analyze several author-identified text collections to compare the statistical quantities. The ratio of recycling is also measured for each collection. Finally, related research topics are introduced together with some discussion of future research directions.

1 Introduction

In conventional information retrieval studies, the similarity between two documents is calculated based on the distribution of terms that appear in each document. However, in document databases, or on the Web, there exist numbers of documents that literally contain the same phrases. These documents not only maintain a good statistical resemblance but also share a long section of terms, sometimes spread over sentences.

When the degree of the match is beyond the level of a simple coincidence, it is a natural consequence that these sections of terms are duplicated and reused by the authors. Furthermore, we can assume that, in this digital age, this type of 'recycling' is an ordinal practice when authoring text-based products because texts are easily copied and reused. Another important aspect is that the reused texts are often semantically meaningful; their survival across documents itself is an evidence of their usefulness. For example, some expressions contain the definitions of named entities that are shared between the two documents.

It should be emphasized here that the statistical similarity and the term sequence matching are strongly associated, but essentially different, phenomena. The former is derived from the topical relationship between the two documents, whereas the author's editing, revising, or quoting a document, indicating some form of 'social' relatedness, causes the latter. However, there have been few attempts, to date, to analyze text corpora explicitly focusing on the reuse and reusability issues.

Based on the above observations, this paper aims at establishing a methodological basis for extracting featured term sequences reused in a group of documents. First, we define the following three types that correspond to distinctive reuse patterns of term sequences.

(1) *Compounds and phrases.* Permanent lexicon and idiomatic expressions that are frequently and universally used in texts.

(2) *Instantly lexiconized texts.* Passages and conventional expressions that are only temporarily and locally reused. Reusable without credits to the authors, also referred to as *instant lexicon.*

(3) *Quoted texts.* Passages that are attributed to a particular author. When used by other authors, usually copied with credits, also referred to as *authored texts.*

Note that we consider only the designated 'writer' of the target text here. Issues in identifying a copyright holder of a specific text are outside of the scope of this paper.

While terms and compounds have long been a central issue of natural language processing studies, little attention has been paid to the extraction and utilization of longer passages, namely, the *instant lexicon* and the *authored texts* as previously defined. Nevertheless, these are the featured text elements that are most strongly related to particular topics or authors, and therefore could be useful resources in various text processing applications, such as authorship identification, duplication checking, document clustering and summarization.

Because the exploration in this direction has just started, in this paper we limit our focus to the following three issues. First, in section **2**, we present an efficient method for extracting reused term sequences together with the corresponding document subsets. Special attention is paid to make the method simple and general so that it is easily applicable to wide variety of text resources. Next, in section **3**, some analytical results are reported where the proposed method was applied to several text collections and the statistical natures were compared. Finally, in section **4**, we introduce related research topics and discuss the utilization of the proposed method in connection with existing text retrieval applications.

2 Suffix Tree based Clustering

2.1 Definition of ST-Clusters

Denote all the documents in the target corpus as $\mathcal{D}$, all the terms in the target corpus as $\mathcal{W}$. The *word n-gram* ($n \geq 1$) is a sequence of n terms given by:

$$w_1^n = (w_1, \cdots, w_n) \qquad (w_i \in \mathcal{W}) \ . \qquad (1)$$

Next, consider a *suffix tree*, each node of which corresponds to a distinctive word n-gram observed in $\mathcal{D}$. For every node on the tree, there exists an uniquely determined subset $S_d(w_1^n)$ ($\subset \mathcal{D}$) given as a subset of all the documents that contain w_1^n. In other words, w_1^n is a sequence of terms that is shared between $S_d(w_1^n)$. Noting that multiple nodes may refer to the same document subset, we define a *suffix tree based cluster* (ST-cluster) as a subset of nodes on the suffix tree that is mapped to the same document set. Namely,

(definition) *A ST-cluster is defined as a pair (S, D) such that S is a subset of n-grams, D is a subset of documents, and $\forall s \subset S$, $S_d(s) = D$, $\forall s \not\subset S$, $s_d(s) \neq D$.*

For example, in Figure 1, nodes A and B both refer to subset { DOC#10, DOC#13 } and are therefore merged into a single ST-cluster.

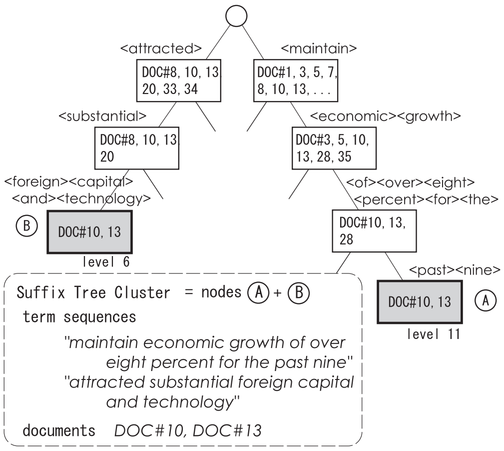


Figure 1: Example of a ST-cluster

2.2 Procedure for ST-Clustering

The basic procedure for extracting ST-clusters is similar to that used by Zamir & Etzioni (1998) and is summarized as follows:

(1) Convert the target collection into sequences of terms, each of which corresponds to a single document. Apply morphological analysis or other pre-processing methods when it is necessary to determine the word boundaries. Neither stemming nor normalization is applied.

(2) Generate a suffix array by a single sort. Suffix tree nodes, together with their corresponding document subset lists, are then

identified as adjacent members of the suffix array. For each node, sort the document list according to some pre-determined order.

(3) Sort all the suffix tree nodes using the sorted document list as a key. Then, the adjacent members of the node list with the same key constitute a single ST-cluster.

The computation time of the above procedure is basically determined by the sort operation in step (3) ($O\,(n\; log\; (n))$), and would be feasible with the power of today's computers. What we found more problematic is the cost of memory to store all the suffix tree nodes and the corresponding document lists at step (2). Figure 2 shows the count statistics for different levels of the suffix tree generated from the Reuters collection, which is also used in our later experiments. Based on the figure, it becomes clear that short length n-grams are the most memory consuming. Because our focus is restricted to longer n-grams, in this paper we consider only the suffix tree nodes with longer than four term sequences.

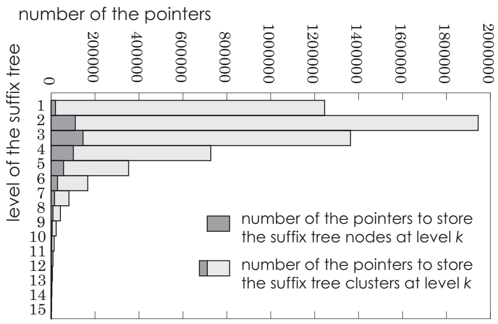


Figure 2: Numbers of pointers at level k

2.3 Measures for ST-Clusters

The ST-clusters generated are evaluated using the following two measures. First is the *term sequence coincidence* that quantifies the strength of the coincidence of the extracted term sequences. Second is the *term distribution similarity* that calculates the divergence of the documents in the cluster based on the conventional document similarity measure.

(1) Term sequence coincidence

The coincidence score of term sequence w_1^n is calculated as the specific mutual information $M(w_1^n)$ given, by definition, as:

$$M(w_1^n) = log\frac{P(w_1^n)}{P(w_1)\cdots P(w_k)}\ . \qquad (2)$$

That is, $M(w_1^n)$ is the difference between (i) the entropy calculated based on the assumption that the k terms $(w_1, \cdots, w_k)$ occurred independently, and (ii) the entropy calculated based on the actual observation.

Intuitively, $M(w_1^n)$ becomes greater for longer sequences. However, the scheme is different from simply counting the length of the sequence because it puts more weight on low frequency terms. In our preliminary experiments, we compared two different rankings of the ST-clusters using $M(w_1^n)$ and the sequence length, and observed the former has a better correlation with the term distribution similarity.

The occurrence probability $P(w_1^n)$ in Eq. (2) is simply determined by $freq(w_1^n)$, the frequency of w_1^n in $\mathcal{D}$, and the overall total frequency F, given as $F = \sum_{w_i \in \mathcal{W}} freq(w_i)$, as follows:

$$P(w_1^n) = \frac{freq(w_1^n)}{F}\ . \qquad (3)$$

The occurrence probability of w_i is also determined by Eq. (3), considering that w_i is a unit length sequence of terms. Because probability estimation of unobserved terms is not an issue here, we have not applied any discounting or smoothing methods for simplicity, unlike many language-modeling studies.

The coincidence score is calculated for every term sequence in the ST-cluster, and then, either the maximum or the total value is used as an overall evaluation, depending on the purpose of the analysis. In this paper, we consistently use the maximum values.

(2) Term distribution similarity

The document similarity of the ST-cluster is defined using the cosine similarity commonly used in information retrieval studies. For each document d in the cluster, index terms are first extracted by applying standard methods, such as morphological analysis, stemming and stop word removal. Then, the term vector $\vec{d}$ is generated for each document using *tf-idf* weighting

Table 1: Data source used in the experiments

Data source	Period	Lang	#Docs	#ST cluster	Execution time	$M(w_1^n)$ per sent.	#Words per sent.
Reuters	1996.8.20 – 1997.8.19	Eng	109,433	1,338,735	2644 sec.	330	24.5
San Jose Mercury	1991.1.1 – 1991.12.31	Eng	72,947	320,457	595 sec.	361	30.1
Mainichi	1998.1.1 – 1998.12.31	Jpn	10,855	111,406	78 sec.	394	33.6
Nikkei	1996.1.1 – 1996.12.31	Jpn	911	19,745	10 sec.	274	27.5
ntc-IPSJ	1988.5.19 – 1997.7.25	Jpn	26,796	226,640	99 sec.	420	32.4
ntc-JSCE	1991.9.17 – 1996.9.17	Jpn	21,259	180,538	70 sec.	434	35.3

scheme. In addition, the central vector of the cluster, denoted as $\vec{c}$, is calculated as an average of all the term vectors. Next, the cosine similarities between the central and each term vector are obtained. Finally, the averaged pairwise similarity values becomes the overall evaluation of the term distribution similarity of the ST-cluster:

$$Sim(D) = \sum_{d \in D} \frac{\vec{d} \cdot \vec{c}}{|\vec{d}||\vec{c}|} . \quad (4)$$

Note that $0 \leq Sim(D) \leq 1$, and the value becomes closer to one for the ST-cluster where the documents are statistically similar to each other.

3 Experiments

3.1 Target Corpora

The six text collections used in our experiments are shown in Table 1. We used two sets of English newspaper articles extracted from either Reuters (Reuters, 2000) or San Jose Mercury (SJM) (Harman & Mark, 1993), two sets of Japanese newspaper articles extracted from either Mainichi (Mainichi, 2001) or NIKKEI (Nikkei, 2001), and two sets of Japanese academic papers' abstracts both extracted from NTCIR-1 (NTCIR, 2001), one presented at the Information Processing Society in Japan (ntc-IPSJ), and the other at the Japan Society of Civil Engineers (ntc-JSCE). For the newspaper articles, we selected only articles with their authors specified, either in the 'byline' (in the case of Reuters and SJM) or embedded in the text in a particular form (in the case of Mainichi and NIKKEI). Morphological analyzer ChaSen was used for Japanese text (Matsumoto, 2001).

For each collection, ST-clusters with term sequences longer than four were enumerated using the method described in **3.2**. The numbers of resulting clusters and the corresponding execution time measured on 2.8GMHz Xeon/Linux are also shown in Table 1. (Note that for comparison purpose, the execution time does not include the time for morphological analysis and word dictionary generation.) For reference, we have also segmented the target collections into sentences, and calculated the average coincidence score together with the average number of terms per a sentence.

3.2 Experiment 1: Measuring the instantly lexiconized texts

In our first experiment, we examine the distribution of the coincidence score of term sequences that were reproduced accidentally without referring to the original document.

For this purpose, we first made mixtures of the two sources: (a) Reuters and SJM, (b) Mainichi and Nikkei, and (c) ntc-IPSJ and ntc-JSCE. Next, we applied the ST-clustering to the generated mixtures. Then, ST-clusters that contain documents from both collections were selected and term sequences with the maximum coincidence score were examined. Note that the pairs were arranged so that both collections belong to the same type (i.e., either newspaper stories or academic papers' abstracts) but originating from different publication sources (i.e., different newspaper companies or academic societies). The topical overlap was also kept small, either by choosing collections of different years, in the case of newspaper articles, or by focusing on different academic fields, in the case of papers' abstracts.

Figure 3 shows the normalized histograms (i.e., empirical p.d.f.) of the coincidence score of the three different mixed pairs. The maximum score is shown in Table 2, together with their length in the parenthesis. Also shown in the table is the *95% threshold value*, such that 95% of the extracted ST-clusters have smaller values than the threshold value.

These results show that the coincidence score is reasonably consistent for all three pairs; between 100 $\sim$ 200 at the maximum, and below 50 for the 95% threshold value. Generally, term sequences beyond this value become candidates for the instant lexicon, indicating some topical relatedness between the documents. However, it should be noted that the threshold cannot be used for discriminating relevant and irrelevant documents because documents without a sequence match can still be similar to each other.

Table 2: ST-clusters associated with articles from different corpora

Data Source	Top score	95% th. value
Reuters + SJM	162.6 (9)	38.7 (6)
Mainichi + Nikkei	116.8 (12)	47.4 (7)
ntc-IPSJ + ntc-JSCE	133.8 (12)	45.5 (8)

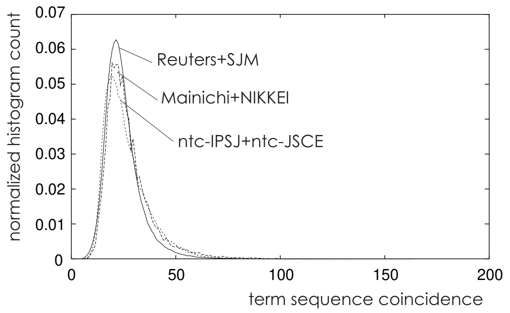


Figure 3: Term sequence coincidence score between irrelevant documents

It is also important to note that the extracted sequences mostly represent semantically meaningful piece of information. The top three examples by Reuters and SJM pairs were (i) "Kate, Larry, Mindy, Nicholas, Odette, Peter, Rose, Sam, Teresa," (score 162.6; a hurricane name list), (ii) "said David Jones, chief economist at Aubrey G. Lanston & Co." (score 124.8), and (iii) "the Golan Heights, which Israel captured from Syria in the 1967 Middle" (score 113.6). We expect that these term sequences are useful in capturing the topical relations between the documents in the same cluster, but this aspect of the ST-cluster is left for future study.

3.3 Experiment 2: Measuring the authored texts

In our second experiment, we grouped the extracted ST-clusters into the following two groups:

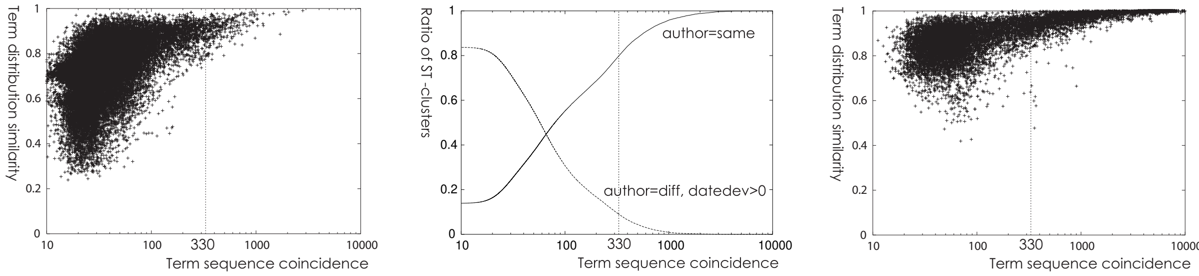
(i) *unique ST-clusters* are composed of articles by the same author (i.e., at least one of the authors is common for all the articles),

(ii) *mixed ST-clusters* are composed of articles by different authors (i.e., none of the authors is common for all the articles).

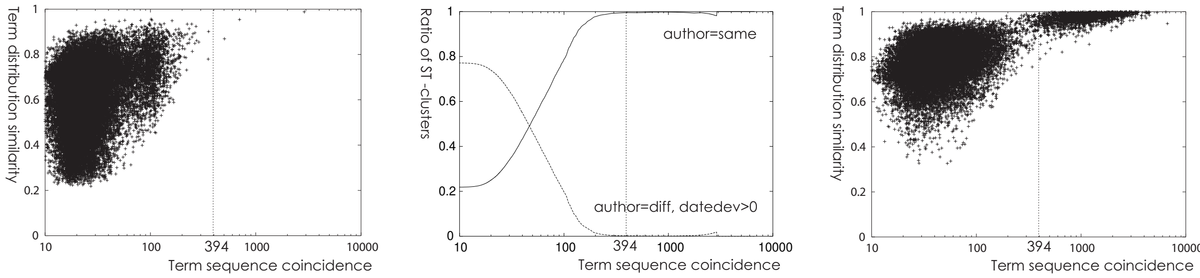
Figure 4 is the result for Reuters, Mainichi, and ntc-IPSJ. The left and the right columns show the relationship between the term distribution similarity and the term sequence coincidence for the mixed and unique ST-clusters respectively, where each point corresponds to a distinctive cluster. The middle column shows the ratio of the two types of the ST-clusters against a different coincidence score. For reference, the average score of a single sentence is also shown as a dotted vertical line, motivated by a naive heuristic that a whole sentence is unlikely to be repeated by chance. For the purpose of readability, only a month's statistics, January 1997, is shown for Reuters (due to the large size). In addition, mixed ST-clusters of documents with the same date were excluded because we found this case contains many miss-identifications, specifically with a series of academic papers co-authored by many researchers.

Based on the Figure 4, it becomes clear that the three collections behave differently in terms of the author structure of the texts. With Reuters, the distinction between the mixed and unique ST-clusters is not obvious. Considerable numbers of articles share exactly the same sentences even when their designated reporters are different. Correspondingly, the changes of the two ratio curves become slow with Reuters.

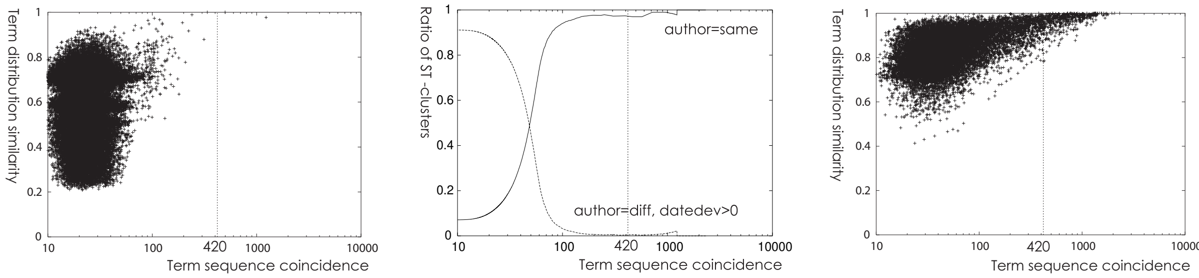
On the other hand, with Mainichi and ntc-IPSJ, there exist only a few ST-mixed clusters



(a) Reuters: mixed ST-clusters (b) Reuters: change on the ratio (c) Reuters: unique ST-clusters



(d) Mainichi: mixed ST-clusters (e) Mainichi: change on the ratio (f) Mainichi: unique ST-clusters



(g) ntc-IPSJ: mixed ST-clusters (h) ntc-IPSJ: change on the ratio (i) ntc-IPSJ: unique ST-clusters

Figure 4: Analysis of ST-clusters associated with articles by multiple or unique authors

with high coincidence scores. Accordingly, the ratio curves for these collections become steeper. In addition, with the Mainichi corpus, the ST-unique clusters are divided into two groups on the graph. On further examination, the clusters in the upper-right region were found to contain different local editions (Tokyo and Osaka) of the same overseas stories sent by the same reporters. Except for these particular cases, the coincidence score of the ST-unique cluster with the Mainichi collection is relatively low compared with the other two collections. For the ntc-IPSJ, exceptional cases were found where a series of papers were presented on the same day but in the name of different authors. However, these cases are excluded from the figure, as have been already described, and cannot be seen in the figure.

The 95% threshold value of the mixed ST-clusters was 907.4 (84) for Reuters, 164.2 (20) for Mainichi, and 110.2 (15) for ntc-IPSJ, where figures in the parenthesis are the length of the term sequences on the border. Compared with the previous case shown in Table 2, the values vary considerably across the different collections. To further clarify the difference, we also examined the influence of the time deviation on the threshold values. This time, we selected only ST-mixed clusters whose time deviation is greater than t, where t varied between $0 \sim 30$ days, and calculated the threshold values for each collection. Based on the result shown in Figure 5, it becomes clear that the text reuse with Reuters is more related to the date of the stories. Only limited length of term sequences were reproduced after an interval of several days.

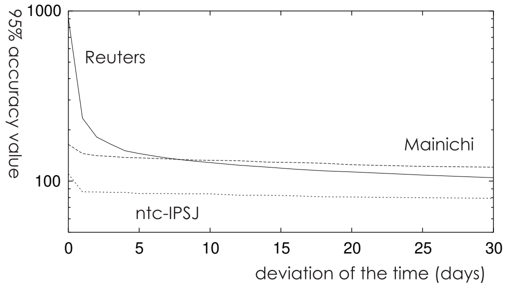


Figure 5: Influence of time deviation

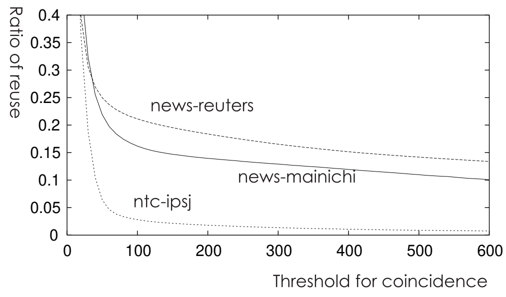


Figure 6: Ratio of text reuse

In summary, the reuse pattern of the authored texts varies across different media and publication styles. With Reuters, the authored texts are generally more likely to be associated with a specific event that occurs within a short period of time. On the other hand, with Mainichi, the texts are instead connected to individual articles of the day, while with ntc-IPSJ, they are simply associated with individual authors or author groups.

3.4 Experiment 3: Measuring the reuse ratio

In our final experiment, we measured the degree of 'recycling' with Reuters, Mainichi, and ntc-IPSJ. We calculated the ratio of term sequences that appeared for more than a second time in the collection, with their coincidence score being greater than a given threshold c. (Their first appearance was not counted in the number.) The value of c was varied $10 \leq c \leq 600$. Figure 6 shows the result. The result shows that the reuse ratio rapidly decreases for c smaller than 50, and then becomes flat for c greater than 100. We noted that the changes correspond approximately to the borders of the instant lexicon and authored texts in the previous experiments, but the details are left for future investigation.

Finally, although the purpose of the experiment is not to compare the reuse ratio of these particular documents, the figures show that the ratio is not negligible for standard collections. The value could be much higher in environments such as the Web.

4 Discussion

Although the analysis of reused text discussed in this paper only captures a particular aspect of text dissemination, related studies are found in several different fields of information retrieval.

(1) Authorship identification

There exist a group of studies concerning the identification of authors (Tony & Michael, 2000). Those stylometric studies quantify the 'style' of a particular author using a combination of various statistical measures, such as the sentence lengths or vocabulary richness. In recent studies, word n-gram information ($N = 3 \sim 10$) is commonly used also. However, past studies mostly focused on extracting discriminators for indicating authors of older, disputed literature archives. It is only recently that the stylometric scheme has been applied to identify anonymous authors (for example, Tsuboi & Matsumoto, 2002). Although the proposed scheme is relevant to copyright issues, the objective is not to detect illegal reprinting intentionally disguised by the authors. Instead, the scheme could be better used to prevent unintentional violation of copyright by inexperienced authors.

(2) Duplicate document detection

A related research topic is duplicate document detection. The topic has become specifically important in recent years due to the explosive increase of documents on the Internet. Chowdhury et al. (2000) categorize the conventional text-based duplicate detection techniques into the following two types: The first is *shingling techniques* where sets of 'shingles', typically contiguous terms, are compared for duplicate detection (Broder et.al., 1997; Chowdhury, et.al., 2002). The second is *similarity measure calculation* where the term distribution similarity is used to detect potential duplicates (Molina, et.al, 1996; Sanderson, 1997). Although most studies allow minor syntactic variations, the duplication is detected for entire documents or Web sites. Because the proposed scheme is focused on partial duplications, it could be used as a complementary measure to improve the flexibility of the duplication check.

(3) Document clustering

There also exist studies that generate clusters based on phrases shared between documents. Suffix Tree Clustering (STC), proposed for on-the-fly reorganization of the search results on the Web, is an example close to our approach (Zamir & Etzioni, 1998). Although both STC and our methods exploit suffix tree structure to realize efficient clustering, the adaptations are slightly different. Because the objective of STC is to create semantically associated document clusters, stemming and sentence level segmentation were applied at the pre-processing stage, term sequences longer than six were penalized with equal weights, and the extracted 'base clusters' are further integrated into larger clusters. Because our focus is on the exact term sequence match, we analyze directly the 'base clusters' extracted from the entire text collections.

Finally, future research directions are as follows. First, the issue of quantifying the authorship of anonymous texts should be further explored, because the interpretation may depend on various factors including the language, the media, the editing policy, or the subject field. The proposed analytical method could be a promising tool to explore different types of textual resources, including Web documents, XML-based databases, or program source codes. The second potential research topic is the rapid detection of partially duplicated texts as well as the automatic generation of embedded text anchors using the proposed clustering method. The third issue concerns the extraction of event-specific expressions that can be utilized further in summarizing the contents of the cluster. The last issue also requires such techniques as re-segmentation and interpolation of terms, and automatic detection of media-specific expressions. In addition, it is important to develop a refined language-based method of identifying and classifying quoted descriptions that appear in the texts.

References

Reuters. Reuters Corpus, Volume 1, English language, 1996-08-20 to 1997-08-19 2000.

Donna Harman and Mark Liberman. TIPSTAR Complete. 1993. Linguistic Data Consortium.

Mainichi Interactive. 1999. 1998 Mainichi Daily News CD-ROM Version.

Nihon Keizai Shinbun. 2001. 1996-2000 Nikkei Full-text Database.

National Center for Science Information Systems. 1999. NTCIR Test Collection 1.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Osamu Imauchi, and Tomoaki Iwamura. 1997. Japanese Morphological Analysis System ChaSen Manual. NAIST Technical Report, NAIST-IS-TR97007.

Tony McEnery and Michael Oakes. 2000. *Authorship Identification and Computational Stylometry.* in Handbook of Natural Language Processing, Marcel Dekker Inc., 545–562.

Yuta Tsuboi and Yuji Matsumoto. 2002. *Authorship Identification for Heterogeneous Documents.* SIG Notes of Information Processing Society in Japan, SIG-NL-148, 17–24.

Abdur Chowdhury, Ophir Frieder, David Grossman, and Mary Catherine McCabe. 2002. *Collection Statistics for Fast Duplicate Document Detection.* ACM Trans. on Information Systems, 20(2), 171–191.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. *Syntactic Clustering of the Web.* Proc. of the Sixth International World Wide Web Conference, 391–404.

Héctor García-Molina, Luis Gravano, and Narayanan Shivakumar. 1996. *dSCAM: Finding Document Copies Across Multiple Databases.* Proc. of Fourth International Conference on Parallel and Distributed Information System, 68–79.

Mark Sanderson. 1997. *Duplicate Detection in the Reuters Collection.* Technical Report of the Department of Computing Science at the University of Glasgow, TR-1997-5.

Oren Zamir and Oren Etzioni. 1998. *Web Document Clustering: A Feasibility Demonstration.* Proc. of SIGIR'98, 46–54.