December 2005

An Evolution-based Approach to Preserving User Preferences in Document-Category Management

Chih-Ping Wei
National Sun Yat-sen Univ.

Paul Hu
University of Utah

Yen-Hsien Lee
National Sun Yat-sen Univ.

An Evolution-based Approach to Preserving User Preferences in Document-Category Management

Chih-Ping Wei
Dept. of Info. Management
National Sun Yat-sen Univ.
Kaohsiung, Taiwan, ROC
cwei@mis.nsysu.edu.tw

Paul Jen-Hwa Hu
Department of Accounting
and Information Systems
University of Utah, USA
actph@business.utah.edu

Yen-Hsien Lee
Dept. of Info. Management
National Sun Yat-sen Univ.
Kaohsiung, Taiwan, ROC
roso@mis.nsysu.edu.tw

Abstract

Document clustering is critical to automated document management, hereby a set of documents are clustered in multiple categories, each containing similar or relevant documents. Most previous research assumes time invariability of document category; i.e., not evolving over time after creation. The adequacy of an existing category understandably may diminish as it includes influxes of new documents over time, bringing about significant changes to its content. Following an evolution-based approach to preserving user preferences in document-category management, this study extends Category Evolution (CE) technique by addressing its inherent limitations. The proposed technique (namely, CE2) automatically re-organizes document categories while taking into account those previously established by the user. We empirically evaluate the effectiveness of CE2 in different document management scenarios that are created using a set of documents from Reuters. Our evaluation includes both CE and hierarchical agglomerative clustering (HAC) for performance benchmarks. Our analysis results show CE2 to be more effective than CE and HAC, showing higher clustering recall and precision. Our findings have interesting implications to research and practice, which are discussed together with our future research directions.

Keywords: Document-category management, Category evolution, Document clustering, Hierarchical clustering analysis

1. Introduction

The rapidly expanding corpus of online documents favors automated document management which can be effectively supported by text-mining based techniques. Document clustering is critical to automated document management, hereby a set of documents are partitioned into multiple categories, each containing a set of documents pertinent to the same or similar topics. Most previous research explicitly or implicitly assumes time invariability of document categories; i.e., not evolving over time after creation. However, the adequacy of a previously created document category can diminish as the category includes influxes of new documents over time, bringing about significant changes to its content and cohesiveness.

A handful of studies (Boley et al., 1999; El-Hamdouchi and Willett, 1986; Larsen and Aone, 1999) took a discovery-based (or rediscovery-based, to be more specific) approach to managing the dynamically evolving document categories. According to this approach, both newly arrived and existing documents are used to create or re-create document categories, without considering the categories previously established. Such discovery-based methods for automated document management may not be effective, particularly with respect to preserving the user's preferences or perspective on the semantic coherence between or among

different documents (Rucker and Polanco, 1997). Preservation of the users' document-grouping preferences indeed represents a fundamental challenge in automated document management. Towards this, Wei et al. (2002) proposed Category Evolution (CE), an evolution-based approach for automated document re-organizations by taking into account the categories previously created. Preliminary evaluations suggest promising capability of CE for retaining the user's preferential perspective in document grouping, showing a clustering accuracy better than that achieved by prevailing discovery-based techniques (Wei et al., 2002).

While appealing, CE has several inherent limitations. First, CE uses intra-category disjointness to evaluate whether or not to decompose a document category, may not be effective across different scenarios. For instance, CE cannot yield desirable category decomposition when a target category evolves multiple subcategories that embrace distinct dominant features but have highly comparable collective feature sets. CE is also constrained by its measurement for assessing an optimal number of document clusters to generate from a document category. Specifically, CE uses the silhouette measurement which depends on the relative distance between subcategories to decide whether a subcategory should be separated (split) from the remaining subcategories. When a subcategory is insignificantly distance from the others, CE might be ineffective in decomposing those neighboring subcategories because their distances to each other become marginal, as compared to that of the distant subcategory.

The current research extends CE by addressing its inherent limitations. We propose CE2, which replaces the collective feature-set based intra-category disjointness with document similarity, and uses intra-category cohesiveness for category-decomposition evaluations. In this study, we describe the design and implementation of CE2 and report the results of an empirical evaluation, based on a real-world document set. Our overall objective is to test whether CE2 is more effective for preserving user preferences in document-category management than CE, which has been shown to outperform prevailing discovery-based methods. We included in our evaluation a prevailing clustering technique (i.e., hierarchical agglomerative clustering) for performance benchmark purposes.

The remaining of the paper is organized as follows. In Section 2, we review previous research of automated document management and provide an overview of CE, together with analysis of its inherent limitations. In Section 3, we discuss the design and implementation of the proposed CE2, followed by our empirical evaluation design and important comparative analysis results in Section 4. We conclude this paper in Section 5 with a summary and discussions of our contributions and future research directions.

2. Literature Review and Overview of CE

Document clustering is essential to automated document management. In general, document clustering partitions a set of documents into different clusters (categories or groups), based on the contents of documents. The documents in the resultant clusters exhibit highest similarity to those in the same cluster and share minimal similarity with documents in other clusters. Common clustering algorithms include partitioning-based (Cutting et al., 1992; Boley et al., 1999; Larsen and Aone, 1999), hierarchical (El-Hamdouchi and Willett, 1986; Roussinov and Chen, 1999; Voorhees, 1986), and Kohonen neural network based (Kohonen, 1989; Kohonen, 1995; Lagus et al., 1996; Roussinov and Chen, 1999). A review of extant literature suggests a predominant focus on complete discovery or re-discovery, hereby creating document categories using all available documents (new and previously existing combined) but does not consider the document categories previously established. Evidently, the

discovery-based approach is not designed to preserve user preferences in document-category management. A review of relevant previous research suggests an assumption that a document category, once created, needs not to evolve over time. However, the adequacy or effectiveness of a document category conceivably can diminish as influxes of new documents arrive over time, bringing about considerable changes to the document category's content.

To cope with the evolving nature of document categories over time, Wei et al. (2002) proposed CE to preserve user preferences in document-category management. In a nutshell, CE takes as inputs existing document categories together with their documents, and generates new document categories, each of which contains documents of increasingly similarity or relevance. The design of CE focuses on single-category documents and fundamentally considers document categories as a set rather than a hierarchy. CE addresses the category evolution requirements by generating a new set of categories through re-organizations of existing categories. CE performs document-category re-organizations when influxes of new documents significantly decrease the adequacy or effectiveness of existing document categories.

Category decomposition and amalgamation are critical to CE. Category decomposition splits an existing document category into multiple new categories, each containing increasingly cohesive documents germane to a topic of finer granularity. In the category decomposition phase, CE first extracts from the documents a set of nouns and noun phrases from which it selects a set of representative features using the TF×IDF selection method for creating a local dictionary for each existing category. Each categorized document is then represented using its particular local dictionary. Intra-category disjointness evaluation is critical to category decomposition because it determines whether or not a document category contains disjointed sets of documents. According to CE, an existing category is tentatively split into two subsets (subcategories), each containing documents that share a greater similarity than those in the other subset. For each existing category, disjointness of the resultant two subsets is assessed using the following intra-category disjointness measurement:

$$disjointness(c, \sigma_d\%) = 1 - \frac{2 \times \left| F_{c_1} \cap F_{c_2} \right|}{\left| F_{c_1} \right| + \left| F_{c_2} \right|}$$

where c is the target category of which documents are clustered into c_1 and c_2 subsets, $\sigma_d\%$ (feature inclusion threshold) is used to eliminate features with low frequency (i.e., less than $\sigma_d\%$ of documents) in the category c, and F_{c_1} (or F_{c_2}) is the set of features each of which appears in at least $\sigma_d\%$ of the documents in c and, at the same time, appears in some document in c_1 (or c_2) subset.

A document category will be decomposed when its intra-category disjointness exceeds a specified threshold (α_s). CE uses the PAM algorithm (Kaufman and Rousseeuw, 1990; Ng and Han, 1994) to decompose a document category into multiple subcategories and then determines an optimal number of categories, based on the silhouette coefficient measure (Kaufman and Rousseeuw, 1990). As a result, all documents in the original category are assigned to appropriate subcategories newly created from the decomposition process.

In the category amalgamation phase, multiple document (sub)categories created from the previous decomposition phase are merged to form a single and more general category that contains documents pertinent to a topic of a broader scope. CE first re-selects features for each new subcategory created previously, using the same feature selection method employed in the decomposition phase. The resulting features are then used to represent individual

documents in the respective document categories or sub-categories. Upon completing feature re-selection and document re-representation, CE evaluates the overlap between document categories. To assess the degree to which two document categories overlap, CE uses an inter-category overlap measurement, which is defined as the following.

$$overlap(c_i, c_j, \sigma_o\%) = \frac{2 \times \left| F_{c_i} \cap F_{c_j} \right|}{\left| F_{c_i} \right| + \left| F_{c_j} \right|}$$

where c_i and c_j are the categories under evaluation, σ_o% (feature inclusion threshold) is used to remove features of low frequency (i.e., less than σ_o% of documents) in each category, and F_{c_i} (or F_{c_j}) is the set of features, each of which appears at least in σ_o% of the documents in c_i (or c_j).

When the overlap between two document categories exceeds a specified threshold (α_m), CE performs category coalescence. To assess and reconcile conflicts that may result from such pair-wise merging evaluations, CE uses a graphical method to analyze merging decisions. For instance, category *A* and category *B* are to be merged. So are category *B* and category *C*, but not category *A* and category *C*. In this graphical analysis method, a node represents a document category and a labeled undirected link indicates the respective merging decisions. The overlap between two categories is represented by a link that connects the representative nodes. Thus, the graph captures all merging decisions suggested by the pair-wise inter-category overlap evaluations. For each connected sub-graph with more than one node (which is not a complete graph), its link with the lowest overlap measure is then removed. This process is repeated until all sub-graphs become complete graphs. Upon completing the category amalgamation phase, CE generates a set of (new) categories which in effect have evolved from those previously created and then re-assigns individual documents to appropriate resulting categories accordingly.

Though CE has shown encouraging effectiveness for preserving user preferences in document grouping, it has several inherent limitations that need to be addressed. First, CE uses intra-category disjointness to evaluate whether or not to decompose a document category. The disjointness is assessed using a collective feature set obtained from all the documents in a category. When assessing plausible decompositions of a document category, CE tentatively splits the category into two subcategories and then calculates their disjointness, based on the respective collective feature sets. CE will proceed with the decomposition when disjointness exceeds a specified threshold. The effectiveness of intra-category disjointness may be constrained in some scenarios. For instance, CE may not be effective in situations where there exist multiple subcategories that embrace distinct dominant features but have highly comparable collective feature sets. As illustrated in Figure 1, the dominant features of the documents in cluster 1 are distinctly different from those of the documents in cluster 2. After tentatively splitting an existing (or original) category into cluster 1 and 2, CE is not likely to proceed with the decomposition because of the great similarity between the respective collective feature sets of cluster 1 and 2. Similar problems might arise in CE's deciding on whether or not to merge two or more document categories. The inter-category overlap, which is also computed according to the respective collective feature sets, might not effectively measure the similarity between or among categories. In turn, this will result in inappropriate amalgamation of document categories. Our analysis suggests such problems are in part attributed to the use of collective feature-set comparisons in intra-category disjointness and inter-category overlap evaluations.

Clusters	Documents	Features									
		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
C_1	d_1	y	y	y	-	y	y	-	-	-	y
	d_2	y	-	-	-	y	y	y	-	-	-
	d_3	-	-	y	-	y	-	-	y	-	-
	d_4	y	-	y	y	-	-	-	-	-	-
	d_5	y	-	-	y	-	y	-	-	-	-
	d_6	-	-	-	y	y	y	-	-	y	-
	d_7	-	-	y	y	y	-	-	-	-	y
	d_8	-	-	y	-	y	y	-	y	-	-
C_2	d_9	-	-	-	-	y	-	y	-	y	y
	d_{10}	-	y	-	-	-	-	-	y	y	y
	d_{11}	y	-	-	-	-	-	y	y	-	-
	d_{12}	-	-	-	-	-	-	y		y	y
	d_{13}	-	-	y	-	-	-	-	y	y	-
	d_{14}	-	-	-	-	-	-	y	-	y	-
	d_{15}	y	-	y	-	-	-	y	-	y	y
	d_{16}	-	y	-	-	y	-	y	y	-	-

Figure 1: Problems of Intra-category Disjointness in Category Decomposition – An Example

The measurement used by CE to determine an optimal number of document clusters represents another important limitation. In the category decomposition phase, CE uses silhouette to determine the number of subcategories to be created from a document category under evaluation. Essentially, silhouette depends on the relative distance between subcategories for deciding whether a subcategory should be separated (split) from the other subcategories. CE may become ineffective in generating subcategories from an existing (original) category. One example is when a subcategory is significantly distant from the other subcategories. As illustrated in Figure 2, subcategory 1 and 2 will be created from an existing category that in effect should be decomposed into 3 document subcategories.

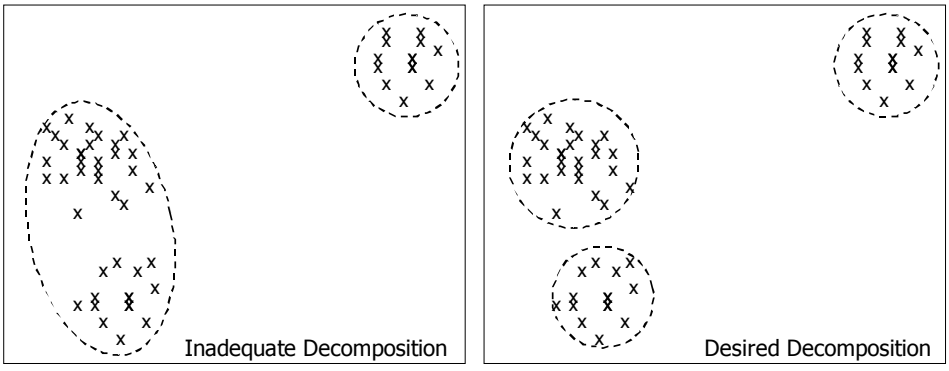

Figure 2: Problem Resulting from Use of Silhouette Measure – An Example

3. Design and Implementation of CE2

To address the described limitations, we propose CE2, which replaces the feature-set based intra-category disjointness with document-based category cohesion, a measurement increasingly effective for identifying inadequacy of existing document categories. Category cohesion is measured by the average similarity of all pairs of documents in a category and therefore is more effective for assessing the appropriateness of document grouping. In addition, CE2 mediates the CE's inherent limitation in category decomposition by

distinguishing most dissimilar documents from other documents in a category and then decomposes the category accordingly. This process is applied to all document categories and subcategories until the cohesion of each (sub)category exceeds a specified threshold.

Figure 3 depicts the overall process of CE2, which essentially performs category decomposition and category amalgamation. In the category decomposition phase, CE2 splits each existing category, if appropriate, into multiple subcategories each of which contains similar documents pertinent to a topic of finer granularity. In the category amalgamation phase, CE2 merges multiple categories or subcategories into a more general category, which contains similar documents on a topic of a broader scope. The detailed design of each phase is as follows.

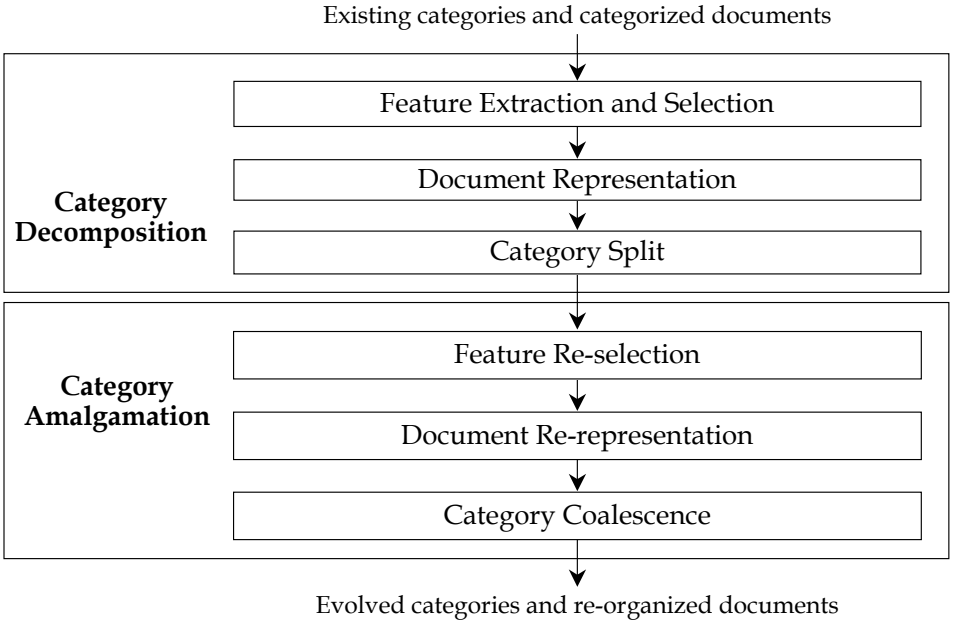

Figure 3: Overall Process of CE2

3.1 Category Decomposition Phase

As shown in Figure 3, the major tasks in the category decomposition phase include *feature extraction and selection*, *document representation*, and *category split*. Contrary to CE, CE2 does not perform intra-category disjointness evaluations; rather, it embeds such evaluations in category split directly. When performing the feature extraction and selection tasks, CE2 extracts from the categorized documents a set of representative features (which consist of nouns and noun phrases) for each existing document category. We use a rule-based part-of-speech tagger technique to syntactically tag each word in a document (Brill, 1992 and 1994). A noun-phrase parser is then applied to extract nouns or noun phrases from each tagged document (Voutilainen, 1993). Subsequently, a set of k_s features is selected for each existing document category on the basis of the TF×IDF feature selection method. Subsequently, we adopt the binary scheme for document representation, which has been shown by previous empirical studies to be capable of producing clustering quality comparable with, if not more favorable to that attained by other schemes (Roussinov and Chen, 1999). Thus, the documents in each existing category are represented using the feature set specific to that document category.

When performing the category split task, CE2 employs the hierarchical divisive clustering method to decompose an existing category into a set of subcategories. Choice of the

hierarchical divisive clustering method over other clustering algorithms (e.g., partitioning-based and Kohonen neural network techniques) is made primarily because it does not require an explicit specification of the number of clusters; thus, can increase or decrease by moving up and down in the resultant clustering hierarchy. For each existing document category, the hierarchical divisive clustering algorithm (Kaufman & Rousseeuw, 1990) starts with all documents in one cluster, and then subdivides the category into two smaller clusters until the average document similarity in every cluster exceeds a predefined similarity threshold (α_s). Imaginably, a lower α_s results in a fewer number of subcategories decomposed from a document category. In this study, the similarity between two documents d_i and d_j is estimated by the cosine similarity measure:

$$sim(d_i, d_j) = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}| \times |\vec{d_j}|}$$

where $\vec{d_i}$ is the feature vector of document d_i and $|\vec{d_i}|$ is the length of $\vec{d_i}$.

3.2 Category Amalgamation Phase

The categories (or subcategories) created in the category decomposition phase become inputs to the category amalgamation phase, in which CE2 merges multiple categories (or subcategories) into more general categories. The major tasks in the category amalgamation phase include *feature re-selection*, *document re-representation*, and *category coalescence*. CE2 first re-performs feature selection across all the categories (or subcategories) created in the category decomposition phase, thus generating a global dictionary comprising of a universal feature set for all document categories (or subcategories). We revise TF×IDF for feature selection by replacing the IDF value of a feature *f*, IDF(*f*), with $\log_2(nc/n_f)+1$, where *nc* is the total number of categories (or subcategories) input to the category amalgamation phase, and n_f is number of categories (or subcategories) containing the feature *f*. The revised TF×IDF is used to measure the power of a feature *f* for characterizing a document category; i.e., distinguishing it from the other categories. The k_m features with the highest TF×IDF scores are then selected as the global dictionary and used to represent each document. As with the document representation task in the category decomposition phase, CE2 adopts the binary scheme for document representation.

Subsequently, CE2 performs category coalescence to merge similar (sub)categories using inter-category similarity, thus creating more general categories. To avoid cyclic processing, CE2 prohibits direct merging of two subcategories (e.g., C_i and C_j) originating from the same category in the category decomposition phase. That is, C_i and C_j can be merged only when there exists another category C_k such that (1) C_k is not originated from the same category as C_i and C_j, and (2) C_i and C_j merge with C_k sequentially rather than simultaneously.

This restriction makes the use of an extended hierarchical agglomerative clustering (HAC) algorithm (Voorhees, 1986) more viable and appealing than other clustering algorithms. The extended HAC algorithm starts with as many clusters as there are categories or subcategories (generated in the category decomposition phase). That is, each document category or subcategory forms a cluster initially. The two clusters exhibiting the highest similarity that exceeds a specified merging threshold (α_m) and not violating the described restriction are then merged to create a new cluster. CE2 uses the group-average link method (i.e., average similarity between all inter-cluster pairs of documents) to measure the similarity between two document clusters. This merging process continues until the similarity of the permissible merges is lower than a pre-specified similarity threshold α_m. Upon the completion of category

coalescence, CE2 generates a set of categories which, in effect, have evolved from the document categories previously created by the user.

4. Evaluation Design and Results

We used the single-category version of Distribution 1.0 of Reuters-21578 document collection[1] to empirically evaluate the effectiveness of CE2. This particular collection of documents has a total of 9,034 single-category documents pertinent to 64 different categories (e.g., topics). We randomly selected 9 categories (i.e., acq, coffee, crude, earn, interest, money-fx, money-supply, sugar, and trade) from the source categories, each having a minimum of 90 documents. Among the chosen categories, two of them (i.e., acq and earn) have a significantly larger number of documents than others (i.e., 2,125 and 3,735 documents, respectively). To maintain a comparable size among the categories evaluated, we randomly selected from these two large categories documents having length between 10 to 30 lines. As a result, our evaluation included a total of 9 categories that collectively have 2,116 documents, each of which has an average of 193 words.

4.1 Evaluation Procedure

We considered the categories specified in Reuters-21578 to be accurate (i.e., true categories). We randomly selected some documents from a category and re-assigned them to another or other categories, thus creating inaccurate document categories that simulate influxes of new documents arriving in existing document categories. Following a specific Gaussian probability distribution, we first decomposed each true category into a dominant subset and multiple minor subsets. Table 1 summarizes the particular category-evolution scenarios to be evaluated, where the number of minor subsets under investigation ranged from 2 to 5 (i.e., from the *Gaussian-3* to *Gaussian-6* distributions). For example, in the *Gaussian-3* distribution, a true category was decomposed into a dominant subset and two minor subsets, which contained 86.6%, 13.1%, and 0.3% of the documents in the true category, respectively. *Gaussian-6* seems to be a reasonable upper bound of the deteriorated document categories because new documents are included in an existing document category over time and therefore its quality (cohesiveness) is not likely to decrease in a rapid and drastic fashion.

Table 1: Evaluation Scenarios – by Gaussian Distributions

Scenario	Dominant	Minor-1	Minor-2	Minor-3	Minor-4	Minor-5
Gaussian-3	86.6	13.1	0.3			
Gaussian-4	68.2	27.2	4.3	0.3		
Gaussian-5	54.7	31.9	10.9	2.2	0.3	
Gaussian-6	45.1	31.8	15.8	5.6	1.4	0.3

For each evaluation scenario, all dominant subsets remained in their respective (true) categories, while each minor subset was randomly merged with the dominant subset from another true category. That is, each minor subset was combined with the dominant subset of a different document category. For each evaluation scenario, we created a synthetic dataset containing a total of 9 inadequate document categories to be clustered by CE2, CE and HAC, respectively. To minimize potential biases resulting from the randomization process when generating a synthetic dataset, we randomly sampled 80% of the documents from the 9 true categories to create a synthetic dataset for a specific evaluation scenario and repeated performed the described process 30 times. The overall effectiveness of each investigated technique using its average performances across the 30 random trials.

[1] Available at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.tar.gz.

4.2 Evaluation Criteria

We measured the effectiveness of CE2 as well as the benchmark CE and HAC using cluster recall and cluster precision, both of which are based on the association of a document pair belonging to the same cluster (Roussinov and Chen, 1999). Cluster recall (*CR*) is calculated as $CR = \frac{|CA|}{|T|}$, where T is the set of associations in the underlying true categories, and CA is the collective set of correct associations contained in both clusters generated by an investigated technique and the true categories. On the other hand, cluster precision (*CP*) is calculated as $CP = \frac{|CA|}{|G|}$, where G denotes the set of associations in the clusters created by an investigated technique. Assume that documents d_1, d_2, …, and d_7 are from two true categories, T_1 and T_2, where $T_1 = \{d_1, d_2, d_3\}$ and $T_2 = \{d_4, d_5, d_6, d_7\}$. In this case, the set of associations in the true categories is T, which consists of $\{(d_1\text{-}d_2), (d_1\text{-}d_3), (d_2\text{-}d_3), (d_4\text{-}d_5), (d_4\text{-}d_6), (d_4\text{-}d_7), (d_5\text{-}d_6), (d_5\text{-}d_7), (d_6\text{-}d_7)\}$. Let the clusters generated by an investigated technique be G_1, G_2, and G_3, where $G_1 = \{d_1, d_2\}$, $G_2 = \{d_3, d_4, d_5\}$, and $G_3 = \{d_6, d_7\}$. The set of associations in the clusters generated by the technique under discussion then is $G = \{(d_1\text{-}d_2), (d_3\text{-}d_4), (d_3\text{-}d_5), (d_4\text{-}d_5), (d_6\text{-}d_7)\}$. As a result, the set of correct associations is CA, which consists of $\{(d_1\text{-}d_2), (d_4\text{-}d_5), (d_6\text{-}d_7)\}$. In this example, $CR = \frac{|CA|}{|T|} = \frac{3}{9} = 0.33$, and $CP = \frac{|CA|}{|G|} = \frac{3}{5} = 0.60$.

To assess the inevitable tradeoff between cluster precision and cluster recall, precision/recall trade-off (PRT) curves were employed. A PRT curve represents the effectiveness of an investigated technique with different merging thresholds; i.e., inter-cluster similarity threshold for HAC and category coalescence merging threshold for both CE and CE2. In this study, we examined the merging threshold for each technique from 0 to 1, in increments of 0.02. Evidently, PRT curves closer to the upper-right corner are more desirable than those closer to the point of origin.

4.3 Evaluation Results and Discussions

Prior to our comparative evaluations, we took a computational approach to tune parameters critical to each investigated technique. Key parameter tuning and comparative evaluation results are highlighted as follows.

***Parameter Tuning for HAC*:** For HAC, the number of features (k) is an important parameter that requires tuning. We examined the effect of the number of features (k), ranging from 50 to 200 in increments of 50. As shown in Figure 4, the overall performance of HAC improved when the number of features (k) increased from 50 to 150. When k further increased from 150 to 200, the resulting effectiveness improvement of HAC was marginal. Together, our parameter-tuning results suggested setting k to 200 for HAC.

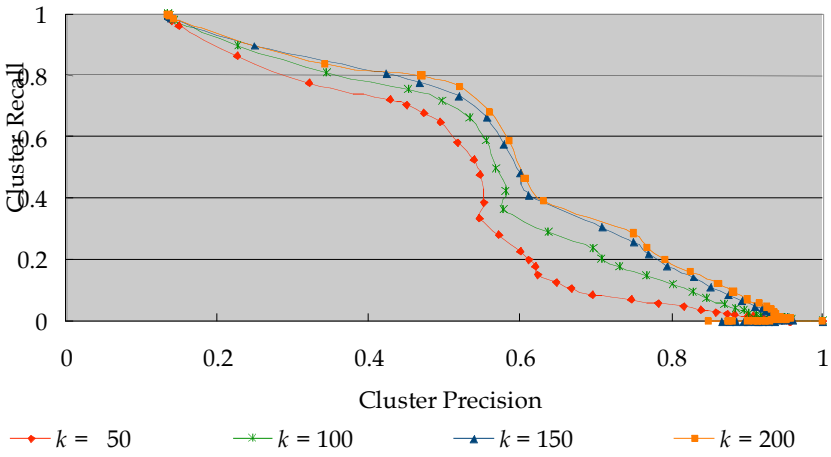

Figure 4: Effects of Numbers of Features for HAC

***Parameter Tuning for CE*:** Key parameters in CE that require tuning include the number of features for category decomposition (k_s), the intra-category disjointness threshold (α_s), and the number of features for category amalgamation (k_m). We first tuned k_s (ranging from 50 to 200 in increments of 50) and α_s (ranging from 0.3 and 0.5 in increments of 0.05) by setting k_m to a default value; i.e., 100. We applied the *Gaussian-4* distribution and followed the evaluation procedure described previously to generate 30 synthetic datasets for parameter tuning purpose. Overall, the resulting CE's effectiveness (as measured by cluster recall and precision) for all k_s–α_s combinations was largely comparable. A better performance was observed when setting k_s at 50 and α_s at 0.3. As a result, we adopted these parameter values in the subsequent experiments.

We then examined different numbers of features for category amalgamation (k_m), ranging from 50 to 200 in increments of 50. According our evaluation results, the impact of k_m appeared to be marginal over the range of k_m value investigated. As a result, we set k_m at 100 with which CE seemed to be relatively effective.

***Parameter Tuning Experiments for CE2*:** CE2 requires the same parameter tuning as CE; i.e., the number of features for category decomposition (k_s), the similarity threshold for category decomposition (α_s), and the number of features for category amalgamation (k_m). The design of the parameter-tuning experiments for CE2 was identical to that for CE. Specifically, we first determined appropriate values for k_s (ranging from 50 to 200 in increments of 50) and α_s (ranging from 0.2 and 0.4 in increments of 0.05) by keeping k_m constant at 200. We used a larger default value than that for CE primarily because CE2 uses a global dictionary for all document categories in the category amalgamation phase as opposed to multiple local dictionaries (used by CE). Thus, it is reasonable to expect CE2 requiring a larger default value for k_m than that by CE.

For increasing interpretability, Figure 5 highlights the parameter value of k_s that yielded the best performance at each value of α_s examined. As shown, we observed k_s appearing to adversely covariate with α_s. That is, given a lower value for α_s (such as 0.2 or 0.25), the use of a higher value for k_s (such as 200 or 150) seemed to produce higher effectiveness by CE2. In contrast, a higher value for α_s (such as 0.35 or 0.4) would require a fewer number of features for category decomposition (such as 50 for k_s). Overall, CE2 seemed to be more effective when k_s = 150 and α_s = 0.25 as well as when k_s = 50 and α_s = 0.35. Accordingly, we adopted 150 for k_s and 0.25 for α_s in the subsequent experiments for evaluating CE2.

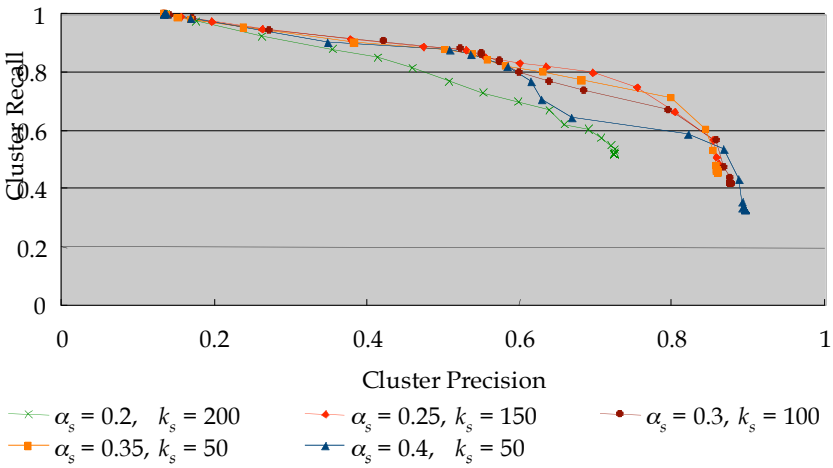

Figure 5: Effects of k_s and α_s for CE2

Based on the selected values for k_s and α_s, we examined effects of numbers of features for category amalgamation (k_m) on CE2's effectiveness. For k_m, we investigated the range between 50 and 200, in increments of 50. As shown in Figure 6, the overall performance improved as k_m increased from 50 to 200. As a result, we set k_m at 200 for CE2.

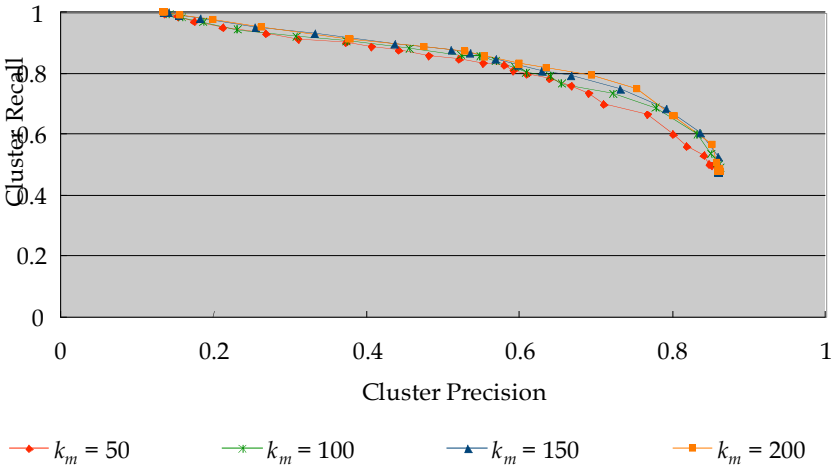

Figure 6: Effects of Numbers of Features for Category Amalgamation (k_m) for CE2

***Comparative Evaluation Results*:** Using the parameter values selected from the described parameter-tuning experiments, we evaluated and compared the effectiveness of CE2, CE, and HAC under different document-management scenarios. As shown in Figure 7-A and 7-B, CE2 and CE significantly outperformed HAC in the *Gaussian-3* and *Gaussian-4* scenarios. Furthermore, the effectiveness of CE2 was noticeably higher than that of CE in these scenarios. As we show in Figure 7-C, CE2 and CE, in the *Gaussian-5* scenario, became less effective when the quality of input document categories deteriorated; nevertheless, both techniques remained advantageous over HAC. The effectiveness of CE appeared to be comparable to that of HAC when the quality of input document categories further deteriorated, and become less effective than HAC in the *Guassian-6* scenario (as shown in Figure 7-D). Overall, our comparative analysis suggests that CE2 was more effective than CE and HAC across the investigated scenarios, and that the effectiveness of CE2 appeared to be more robust than that of CE across the range of input document-category quality examined.

Even in the *Guassian-6* scenario, CE2 was still more effective than HAC, whereas CE was less effective than HAC.

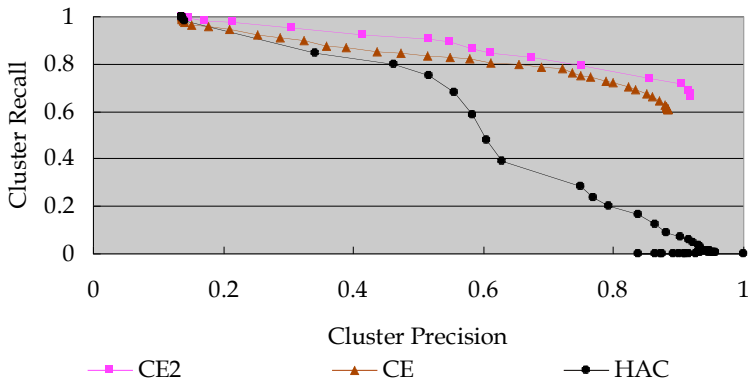

Figure 7-A: Analysis Result – *Gaussian-3* Distribution Scenario

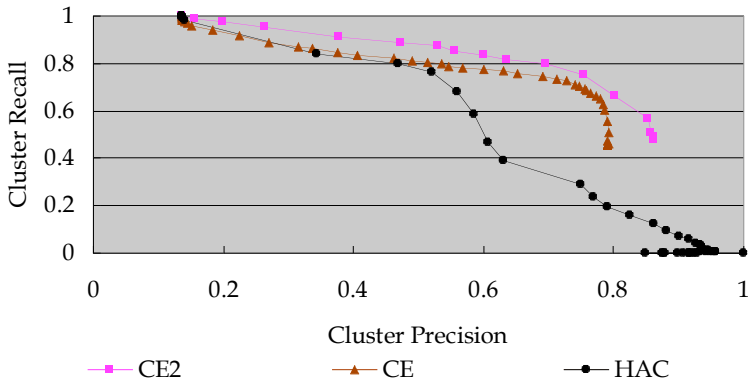

Figure 7-B: Analysis Result – Gaussian-4 Distribution Scenario

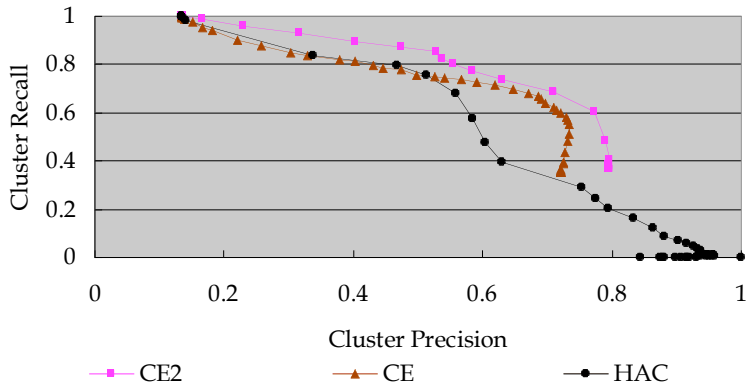

Figure 7-C: Analysis Result – Gaussian-5 Distribution Scenario

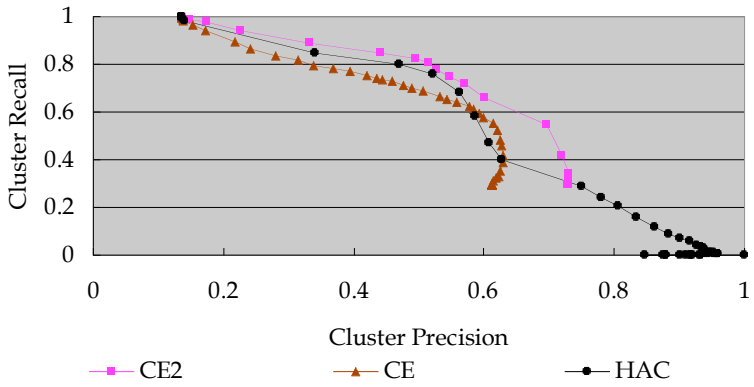

Figure 7-D: Analysis Result – *Gaussian-6* Distribution Scenario

5. Conclusion

We design and implement CE2, which extend from an evolution-based document management technique (CE) by addressing its inherent limitations. Judged by its cluster recall and cluster precision, CE2 exhibits satisfactory effectiveness across different category evolution scenarios. Overall, our evaluation shows CE2 outperforming the benchmark CE and discovery-based technique (i.e., HAC). In addition, the effectiveness of CE2 appears to be reasonably robust with respect to the quality of input document categories.

The study has made several research contributions. First, this research investigates and develops a better text-mining based technique for preserving user preferences in document-category management, which has become increasingly critical in the emerging digital world. This study also contributes to general document management research by responding to the evolving nature of existing document categories, a fundamental challenge that has not yet received due attention by previous research. Results from our comparative evaluations also shed light on the relative value, desirability and limitations the evolution-based and the discovery-based approaches that are critical to document clustering research. While primarily designed for textual documents, the proposed technique can be extended to manage other online resources. Last but not least, our findings can lead to advanced design and evaluation of similar systems in document management or related areas.

This study has several limitations that demand our future research attention. First, our evaluation used simulated rather than real-world scenarios. To mediate this limitation, we are currently designing further evaluations that involve human subjects and use real-world document-management contexts. Second, this study focuses on single-category documents. Understandably, a document may simultaneously pertain to multiple categories (to equal or differential degrees). In turn, this requires effective category management capable of dealing with multi-category documents. In addition, this research concentrates on categories not hierarchically structured; i.e., using a set. Hence, the proposed technique needs to be further extended for multi-category documents and following a hierarchical category structure.

Acknowledgment

This work was supported in part by the MOE Program for Promoting Academic Excellence of Universities of the Republic of China under the grant 91-H-FA08-1-4 and by the National Science Council of the Republic of China under the grant NSC 92-2917-I-110-002.

References

Boley, D., Gini, M., Gross, R., Han, E., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, J. "Partitioning-based Clustering for Web Document Categorization," *Decision Support Systems* (27:3), 1999, pp. 329-341.

Brill, E. "A Simple Rule-based Part of Speech Tagger," *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992, pp. 152-155.

Brill, E. "Some Advances in Rule-based Part of Speech Tagging," *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1994, pp. 722-727.

Cutting, D., Karger, D., Pedersen, J., and Tukey, J. "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections," *Proceedings of 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992, pp. 318-329.

El-Hamdouchi, A. and Willett, P. "Hierarchical Document Clustering Using Ward's Method," *Proceedings of ACM Conference on Research and Development in Information Retrieval*, 1986, pp. 149-156.

Kaufman, L. and Rousseeuw, P. J. (eds.). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, NY, 1990.

Kohonen, T. (ed.). *Self-Organization and Associative Memory*, Springer, 1989.

Kohonen, T. (ed.). *Self-Organizing Maps*, Springer, 1995.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. "Self-organizing Maps of Document Collections: A New Approach to Interactive Exploration," *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996.

Larsen, B. and Aone, C. "Fast and Effective Text Mining Using Linear-time Document Clustering," *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 16-22.

Ng, R. and Han, J. "Efficient and Effective Clustering Methods for Spatial Data Mining," *Proceedings of International Conference on Very Large Data Bases,* Santiago, Chile, Sept. 1994, pp. 144-155.

Roussinov, D. and Chen, H. "Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques," *Decision Support Systems* (27:1), 1999, pp. 67-79.

Rucker, J. and Polanco, M. J. "Siteseer: Personalized Navigation for the Web," *Communications of the ACM* (40:3), March 1997, pp. 73-75.

Voorhees, E. M. "Implementing Agglomerative Hierarchical Clustering Algorithms for Use in Document Retrieval," *Information Processing and Management* (22), 1986, pp. 465-476.

Voutilainen, A. "NPtool: A Detector of English Noun Phrases," *Proceedings of Workshop on Very Large Corpora*, 1993.

Wei, C. and Hu, P. J., and Dong, Y. X. "Managing Document Categories in E-commerce Environments: An Evolution-based Approach," *European Journal of Information Systems* (11:3), 2002, pp. 208-222.