Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2007 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2007

Context-aware Document-clustering Technique

Chin-Sheng Yang
National Sun Yat-sen University, Kaohsiung, Taiwan, litony@mis.nsysu.edu.tw

Chih-Ping Wei
National Tsing Hua University, Hsinchu, Taiwan, cpwei@mx.nthu.edu.tw

95. Context-aware Document-clustering Technique

Chin-Sheng Yang
College of Management,
National Sun Yat-sen University, Kaohsiung,
Taiwan, R.O.C
litony@mis.nsysu.edu.tw

Chih-Ping Wei
College of Technology Management,
National Tsing Hua University, Hsinchu,
Taiwan, R.O.C
cpwei@mx.nthu.edu.tw

Abstract

Document clustering is an intentional act that should reflect individuals' preferences with regard to the semantic coherency or relevant categorization of documents and should conform to the context of a target task under investigation. Thus, effective document-clustering techniques need to take into account a user's categorization context defined by or relevant to the target task under consideration. However, existing document-clustering techniques generally anchor in pure content-based analysis and therefore are not able to facilitate context-aware document-clustering. In response, we propose a Context-Aware document-Clustering (CAC) technique that takes into consideration a user's categorization preference (expressed as a list of anchoring terms) relevant to the context of a target task and subsequently generates a set of document clusters from this specific contextual perspective. Our empirical evaluation results suggest that our proposed CAC technique outperforms the pure content-based document-clustering technique.

Keywords: Document clustering, Context-aware document-clustering, Personalized document-clustering, Text mining, Knowledge management

Introduction

With the advances and proliferation of the Internet, available information sources have grown tremendously in number and sheer volume, primarily as a result of global connectivity and ease of publishing. To manage this ever-increasing volume of documents, organizations and individuals typically organize documents into categories (or category hierarchies) to facilitate their document management and to support subsequent document retrieval and access. In turn, the development of an effective document-clustering mechanism becomes essential to efficient and effective document management of organizations and individuals.

Document clustering entails the automatic organization of a large document collection into distinct groups of similar documents that reflect general themes hidden within the corpus (Kim & Lee, 2000; Kim & Lee, 2002; Pantel & Lin, 2002; Wei et al., 2006a). However, according to the context theory of classification, document-clustering behaviors of individuals not only involve the attributes (including contents) of documents but also depend on who is performing the task and in what context (Barreau, 1991; Case, 1991; Kwasnik, 1991; Lakoff, 1987). As a result, document clustering is an intentional act that should reflect individuals' preferences with regard to the semantic coherency or relevant categorization of documents (Rucker & Polanco, 1997) and should conform to the context of a target task under investigation. That is, when performing a particular task, an individual prefers the categorization of a collection of documents consistent or comparable to the context of the task under consideration. For example, given a set of research articles related to "data mining," an individual who are interested in developing new data mining techniques may prefer a set of document categories anchored at techniques under discussion (e.g., classification analysis, clustering analysis, association rules, and sequential patterns), whereas

the same individual may prefer a different document categories based on application domains involved (e.g., banking, retailing, health care, and telecommunications) when he or she is working on data mining applications. The aforementioned examples highlight the importance of clustering the same set of documents into different document categories for different task contexts concerned by the same individual. Effective document-clustering techniques therefore need to be able to take into account a user's categorization context defined by or relevant to the target task under consideration.

Traditional document-clustering techniques generally anchor in pure content-based analysis. That is, most of existing document-clustering techniques rely on a specific feature selection metric (e.g., term frequency (TF) or TF×IDF (term frequency×inverse document frequency)) (Boley et al., 1999; Larsen & Aone, 1999; Pantel & Lin, 2002; Roussinov & Chen, 1999; Wei et al., 2006a) that are objective in nature to identify a set of representative features as the basis for document clustering. As a consequence, existing document-clustering techniques create a set of clusters that are not tailored to individuals' categorization contexts and therefore are not able to facilitate context-aware document-clustering. The categorization scheme exhibited in such context-unaware clusters may not conform to that of an individual's expectations and perceptions under a specific context. However, an individual's document search typically is guided by his or her own categorization scheme (Donovan, 1991; Restorick, 1986). Thus, when searching documents with a one-for-all categorization scheme, an individual generally undertakes a semantic internalization process (Quillian, 1968) to comprehend the target categorization scheme or experiences a coadaptation process that adjusts his or her own categorization scheme and, at the same time, reinterprets and adapts the target categorization scheme to his or her needs (Mackay, 1988; Mackay, 2000). The semantic internalization and coadaptation processes unnecessarily increase the individual's cognitive load. Consequently, he or she likely spends more time or has difficulty locating documents of interest because of the discrepancy between the one-for-all categorization scheme and his or her expectation (Wei et al., 2006b; Wei et al. 2007). The described inefficiency or ineffectiveness of document retrieval and access may adversely affect the efficiency, quality, and satisfaction of decision making that requires references to various documents relevant to the target decision context.

In response to the limitation of existing document-clustering techniques and the needs of supporting context-aware document-clustering, we propose a Context-Aware document-Clustering (CAC) technique that takes into consideration a user's categorization preference relevant to the context of a target task and subsequently generates a set of document clusters from this specific contextual perspective. The CAC technique assumes that a user's categorization context be expressed as a list of anchoring terms. For instance, given a set of research articles related to "data mining," suppose a user prefers a categorization context from the "application domain" perspective and describes this particular categorization context with such anchoring terms as "banking," "retailing," "health care," "telecommunications," etc. The CAC technique takes as its input the list of anchoring terms provided by the user and attempts to cluster the set of research articles into such various application categories. However, because the list of anchoring terms tends to be small in size and may not contain sufficient information for effectively clustering a target document collection, we incorporate an anchoring term expansion mechanism in our proposed CAC technique. Specifically, the CAC technique exploits World Wide Web (WWW), possibly the largest repository in the world, as the information source for constructing a statistical-based thesaurus and then expands the set of anchoring terms by adding their relevant terms on the basis of this statistical-based thesaurus. Subsequently, the CAC technique uses the expanded set of

anchoring terms for representing the target documents and performs document clustering accordingly.

The remainder of this paper is organized as follows. Section 2 details the existing document-clustering techniques relevant to this study and highlights their limitations in supporting context-aware document-clustering to justify our research motivation. In Section 3, we depict the detailed design of the proposed CAC technique. Subsequently, we depict our experimental design and discuss important evaluation results in Section 4. Finally, we conclude with a summary and some future research directions in Section 5.

Literature Review

In this section, we review the literature on existing document-clustering techniques (including content-based, non-content-based, and hybrid document-clustering ones) and analyze their applicability of and limitations in supporting context-aware document-clustering.

Content-based Document-Clustering Techniques

In essence, document clustering groups similar documents into clusters. The documents in the resultant clusters exhibit maximal similarity to those in the same cluster and, at the same time, share minimal similarity with documents in other clusters. Most of prior document-clustering techniques are anchored in document content analysis. The overall process of a content-based document-clustering technique generally comprises three main phases: feature extraction and selection, document representation, and clustering (Jain et al., 1999; Wei et al., 2002; Wei et al., 2006a). The purpose of feature extraction and selection is to extract and select from the target document corpus a set of representative features (or keywords) to represent the documents in the document representation phase. Subsequently, the clustering phase applies a clustering technique to group the target documents into distinct clusters.

Feature extraction begins with the parsing of each source document to produce a set of nouns and noun phrases (referred to as "features") and exclude a list of prespecified "stop words" that are non-semantic-bearing words. Subsequently, representative features are selected from the set of extracted features. Feature selection is important for clustering efficiency and effectiveness, because it not only condenses the size of the extracted feature set, but also reduces the potential biases embedded in the original (i.e., nontrimmed) feature set (Roussinov & Chen, 1999; Yang & Chute, 1994). Commonly used feature selection metrics include: TF, TF×IDF, and their hybrids (Boley et al., 1999; Larsen & Aone, 1999).

On the basis of a particular feature selection metric, the k features with the highest selection metric scores then are selected to represent each source document in the document representation phase. Based on the chosen representation scheme, each document is then described in the k-dimensional space and represented as a feature vector. Commonly employed document representation schemes include binary (which considers simply the presence or absence of a feature in a document), within-document TF, and TF×IDF (Boley et al., 1999; Larsen & Aone, 1999; Pantel & Lin, 2002; Roussinov & Chen, 1999; Wei et al., 2006a).

In the final phase of document clustering, source documents are grouped into distinct clusters on the basis of the selected features and their respective values in each document. Common clustering approaches include partitioning-based (Boley et al., 1999; Cutting et al., 1992; Larsen & Aone, 1999), hierarchical (El-Hamdouchi & Willett, 1986; Roussinov & Chen, 1999; Talavera & Bejar, 1999; Voorhees, 1986; Wei et al., 2006a), and Kohonen neural

network (Lagus et al., 1996; Lin et al., 1999-2000; Roussinov & Chen, 1999).

As mentioned, content-based document-clustering techniques rely on an objective feature-selection metric (e.g., TF or TF×IDF) that merely considers document content. As a result, existing content-based techniques generate for all users an identical set of document clusters from a given document collection and, thus, is unable to support context-aware document-clustering.

Non-content-based and Hybrid Document-Clustering Approaches

Prior research has proposed non-content-based and hybrid document-clustering approaches (Deogun & Raghavan, 1986; Kim & Lee, 2000) that may be applied to support context-aware document-clustering. For instance, Deogun and Raghavan (1986) propose a user-oriented document-clustering technique, which is solely based on information on document relevance to user queries. Given a document collection $D = \{d_1, d_2, \ldots, d_n\}$ to be clustered and a set of user queries $Q = \{q_1, q_2, \ldots, q_m\}$ (assume that the set of retrieved and relevant documents for each q_j be D_{q_j}), its process consists of two main phases: divisive and merging. In the divisive phase, D is divided into a number of clusters according to document relevance to Q. Initially, D is assumed to form a single cluster. As the first query q_1 is processed, the cluster is divided into two clusters corresponding to relevant and non-relevant sets to q_1. Afterwards, when a new query q_i is processed, each existing cluster that has non-empty intersection with D_{q_j} is divided into two clusters (i.e., relevant and non-relevant sets). The division process continues until all queries in Q are processed. In the merging second phase, the clusters obtained previously are combined to form larger clusters. For each cluster C_i produced in the divisive phase but has not been combined in this phase, its affinity to every other cluster C_j is calculated on the basis of whether their constituent documents co-occur in the set of relevant documents for each query. Accordingly, C_i is combined with the cluster for which their affinity is maximal and satisfies a pre-defined threshold. After the merging phase processes all of the clusters generated in the divisive phase, the document clustering process concludes and the resultant clusters represent the clusters for D.

Alternatively, Kim and Lee (2000) propose a hybrid document-clustering approach to improve clustering effectiveness. Their approach essentially is a semi-supervised technique that considers not only content similarity but also user's perception of document similarity using a relevance-feedback mechanism. Specifically, this hybrid technique consists of three main phases, including preclustering, supervising, and reclustering phases. The preclustering phase, which employs the hierarchical agglomerative clustering (HAC) algorithm (Voorhees, 1986), puts each document in a separate cluster and merges those two clusters whose merger produces the smallest increase in diameter. The merging process then repeats until the diameter of the merged clusters reaches a given threshold. Each of such resultant clusters is referred to as a "precluster." Subsequently, the supervising phase involves obtaining relevance feedback from a user for cluster formation in the later phase. It determines the training document set T that includes all documents within preclusters of less than η documents. Accordingly, a document d_i in T is randomly selected to serve as the query. Using this query, a set of documents in T is retrieved and presented to the user, who then judges whether each of the retrieved documents is relevant to the query (i.e., d_i). Thus, two types of document bundles are formed for d_i: positive and negative. The documents in the positive bundle, which the user has judged as relevant to d_i, are placed in the same cluster as d_i, whereas the documents in its negative bundle must be located in clusters other than d_i. The final reclustering phase involves the formation of clusters for the entire document collection. The preclusters created in the first phase are assigned to the nearest positive bundle. At every

precluster assignment, larger clusters are generated and the set of local cluster prototypes are incrementally updated. Finally, each residual document, which has not been retrieved or has been ignored during the relevance-feedback process, is assigned to the cluster with the nearest local prototype. At this point, documents in negative bundles are examined to check whether they are located in the same clusters. If such documents are found, each of them will be reassigned to the cluster with the document's second nearest local prototype.

To support context-aware document-clustering with the non-content-based approach (i.e., the user-oriented technique), queries employed in the clustering process need to appropriately be selected on the basis of their relevancy to a target categorization context. For example, to organize "data mining" articles from the "application domain" perspective, such queries as "banking," "manufacturing," "health care," and "telecommunications" should be considered relevant to this specific categorization context, whereas such non-application-bearing query terms as "decision tree" and "neural network" are deemed irrelevant. The selection of relevant queries to a target categorization context may not be straightforward and demands a further research attention.

On the other hand, use of the hybrid document-clustering approach (i.e., the semi-supervised technique) to facilitate context-aware document-clustering is less complicated technically than are the non-content-based approach, due to relevance feedbacks of a user involved in its document clustering process. That is, during its supervising phase, if the user has a specific categorization context in his/her mind, he/she would then determine whether each of the retrieved documents is relevant to a query document d_i on the basis of the specific context. Although the hybrid document-clustering approach can support context-aware document-clustering, real-time relevance feedbacks from the user can be time consuming and impractical. This practical limitation becomes even greater as the size of the document collection to be clustered expands.

Context-Aware Document-Clustering (CAC) Technique

We propose the CAC technique in response to the abovementioned limitations of existing document-clustering techniques in supporting context-aware document-clustering. As mentioned, the proposed CAC technique is guided by a user's categorization context represented as a list of anchoring terms and a statistical-based thesaurus constructed by exploiting the World Wide Web (WWW) as the information source and subsequently generates a set of document clusters from this particular preferential context of the user. Figure 1 shows the overall process of the CAC technique, which consists of five main phases: 1) feature extraction and selection; 2) statistical-based thesaurus construction; 3) anchoring term expansion; 4) document representation; and 5) clustering. In the following, we will describe the detailed design of each phase in the CAC technique.

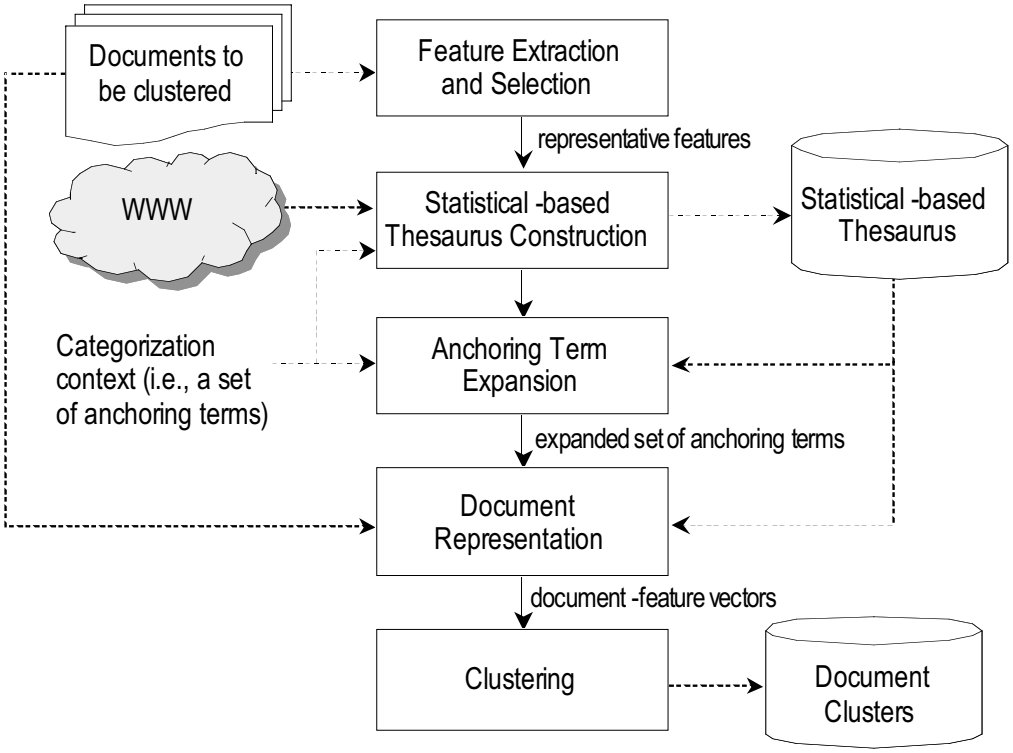

Figure 1: Overall Process of the CAC Technique

Feature Extraction and Selection: The purpose of this phase is to extract and select a set of representative features (specifically, nouns and noun phrases) from the target document corpus (i.e., the collection of documents to be clustered). This set of representative features forms the basis for anchoring term expansion. We adopt the rule-based part-of-speech tagger developed by Brill (1992, 1994) to syntactically tag each word in the target documents. Subsequently, we employ the approach proposed by Voutilainen (1993) to implement a noun-phrase parser for extracting noun phrases from each syntactically tagged document. Furthermore, we remove features that infrequently appear in the target document corpus. Particularly, we only retain those features whose document frequency is no less than a prespecified threshold δ_{DF}.

Statistical-based Thesaurus Construction: The purpose of this phase is to automatically construct a statistical-based thesaurus that will be used for expanding the user-provided anchoring terms relevant to his or her categorization context. CAC exploits the World Wide Web (WWW) to create the statistical-based thesaurus, which will serve as the basis for expanding the set of anchoring terms relevant to the categorization context of a user. Because WWW probably is the largest repository in the world, the association strength (or relevance weight) between two terms measured by the co-occurrence analysis on a search engine's query results will have higher statistical reliability than that estimated from the co-occurrence analysis on a smaller document corpus (Turney & Littman, 2003).

For each anchoring term q_i pertaining to the categorization context of a user and every feature t_k representative to the target document corpus, we issue three queries (i.e., q_i, t_k, and $q_i \wedge t_k$) to a search engine (specifically, Google in this study) and obtain the number of hits (matched documents) returned for each query. We denote the collection of queries for the intended clustering task as a context-aware document-clustering session. The relevance weight

between q_i and t_k is then estimated by the pointwise mutual information (PMI) (Turney & Littman, 2003) as follows:

$$rw_{q_i,t_k} = \log_2\left(\frac{p(q_i \wedge t_k)}{p(q_i)\,p(t_k)}\right) = \log_2\left(\frac{N\times\text{hits}(q_i \wedge t_k)}{\text{hits}(q_i)\,\text{hits}(t_k)}\right)$$

where rw_{q_i,t_k} denotes the relevance weight between q_i to t_k, $p(query)$ is the probability that *query* occurs in the repository (i.e., WWW in our study), N is total number of documents in the repository, and hits(*query*) is the number of hits returned by the search engine of choice. Because the exact value of N in the WWW environment is difficult to estimate, we employ an alternative approach, which sets N as the largest hit value among all the queries in the context-aware document-clustering session we issue to the search engine.

Figure 2 shows an example of the statistical-based thesaurus. For instance, the relevance weights between "data mining" and "clustering analysis," "sequential pattern," "association rule," and "classification analysis" are 3.75, 2.84, 3.28, and 4.34, respectively. However, the relevance weight between "data mining" and "outsourcing," two semantically unrelated terms, is only 0.12.

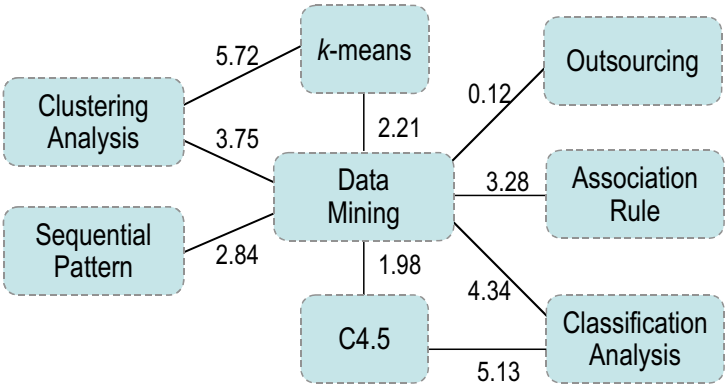

Figure 2: Example of Statistical-based Thesaurus

Anchoring Term Expansion: This phase is to expand the set of anchoring terms AT by including additional relevant terms on the basis of the statistical-based thesaurus constructed previously. An anchoring term q_i in AT is expanded with a set of terms E_{q_i} whose relevant weights to q_i need to be greater than a prespecified threshold α. Take the previously hypothetical thesaurus as an example. If the anchoring term q_i to be expanded is "data mining" and α is 3, the expanded set of terms E_{q_i} will consist of "clustering analysis," "association rule," and "classification rule." Accordingly, the resultant expanded set of anchoring terms $RF = \left(\bigcup_{q_i \in AT} E_{q_i}\right) \cup AT$ is formed for the subsequent document-clustering task.

Because RF consists of the anchoring terms originally provided by the user and relevant terms expanded from the anchoring terms, the importance of the terms in RF should not be identical when they are used to represent each document to be clustered. For example, the original anchoring terms should be more important than any of the expanded terms. In addition, an expanded term with a higher relevance weight to the original anchoring terms should be more important than that with a lower relevant weight. However, if an expanded term is associated with too many anchoring terms, it is possible that the expanded term is too general to discern concepts embedded in the preferential context of the user and topics discussed in documents. In this case, its importance should be reduced. Accordingly, we

adopt the TF×IDF scheme and define the weight of each expanded term f_j in RF but not in AT as:

$$w_j = \sum_{q_i \in AT_j} rw_{q_i,t_k} \times \log\left(\frac{|AT|}{|ET_j|} + \varepsilon\right)$$

where ET_j is the set of anchoring terms that expand f_j and ε is a small positive value to avoid the log component in the formula being 0.

On the other hand, if $f_i \in AT$, w_j is the largest weight across all expanded terms derived previously.

Document Representation: This phase is to represent each document to be clustered using the expanded set of anchoring terms RF. In this study, we employ the TF×IDF scheme weighted by the weight of each term in the expanded set of anchoring terms for document representation. Specifically, each document d_l is described by a feature vector $\vec{d_l}$ as:

$$\vec{d_l} = \langle v_{l1} \times w_1, v_{l2} \times w_2, \ldots, v_{lm} \times w_m \rangle,$$

where m is the total number of terms in RF, v_{lj} is the TF×IDF value of f_j in d_l, and w_j is the weight of the term f_j in RF.

Clustering: In the final phase, the target documents are grouped into distinct clusters on the basis of the expanded set of anchoring terms (i.e., RF) and their respective values in each document. Among the common document-clustering approaches (including partitioning-based, hierarchical, and Kohonen neural network), hierarchical clustering has an advantage over partitioning-based, in that the number of clusters need not be prespecified and can be decreased (or increased) by adjusting the intercluster similarity threshold. Furthermore, the hierarchical clustering approach could achieve clustering effectiveness comparable to the Kohonen neural network (Roussinov & Chen, 1999). Therefore, we adopt the hierarchical clustering approach (specifically, the HAC algorithm) as the underlying clustering algorithm for our proposed CAC technique. In addition, we adopt the cosine measure to estimate the similarity between two documents and employ the group-average link method for measuring the similarity between two clusters. That is, two clusters whose average similarity among all intercluster pairs of documents is the highest will be joined first.

Empirical Evaluation

This section reports our empirical evaluation of the proposed CAC technique using a traditional content-based document-clustering technique (specifically, the HAC algorithm using the TF×IDF feature selection metric and group-average link method) as a performance benchmark. In the following, the evaluation design (including data collections and evaluation criteria), parameter tuning experiments, and important evaluation results will be detailed.

Data Collection

The collection of document corpus for our evaluation purpose consisted of 434 research articles related to information systems and technologies that were collected through keyword searches (e.g., XML, data mining, robotics) from a scientific literature digital library website (i.e., CiteSeer, http://citeseer.nj.nec.com/). For each article in our Literature corpus, only the abstract and keywords were used in this evaluation study.

> To evaluate the effectiveness of the proposed CAC technique, we need to categorize our Literature corpus from different contextual perspectives. We developed a Web-based system to collect individuals' preferred clustering for the Literature corpus. Because the target corpus

relate to information systems and technologies, we constrained our experimental subjects to master and doctoral students majoring in management information systems. Each experimental subject was asked to categorize the randomly ordered documents manually. After clustering, the subject was asked to assign a label for each category. These category labels are then considered as the set of anchoring terms with respect to the categorization context relevant to his or her clustering of the corpus and will be used as the input to the CAC technique. A total of 33 subjects accomplished the manual clustering of the documents in the Literature Corpus. According to the self-reported estimates of the subjects, each subject spent a minimum of eight hours performing manual document clustering. A summary of the document categories generated by the subjects is provided in Table 1.

Table 1: Summary of Subjects' Categories for the Literature Corpus

	Number of Folders	Number of Documents in a Folder
Maximum	67	125
Minimum	10	1
Average	26.12	16.64

Evaluation Criteria

We employ cluster recall and cluster precision (Roussinov & Chen 1999), defined according to the concept of associations, to measure the effectiveness of the CAC technique and its benchmark technique. An association refers to a pair of documents that belong to the same cluster. Accordingly, the cluster recall (*CR*) and cluster precision (*CP*) from the viewpoint of a subject u_a is defined as:

$$CR = \frac{|CA_a|}{|T_a|} \text{ and } CP = \frac{|CA_a|}{|G_a|}$$

where T_a is the set of associations in the categories manually produced by the subject u_a, CA_a is the set of correct associations that exists in both the clusters generated by a document-clustering technique and the categories produced by u_a, and G_a is the set of associations in the clusters generated by the document-clustering technique.

To address the inevitable trade-offs between cluster recall and cluster precision, precision/recall trade-off (PRT) curves are employed. A PRT curve represents the effectiveness of a document-clustering technique with different intercluster similarity thresholds. Evidently, as the intercluster similarity threshold increases, the average number of documents in each cluster decreases; thus, generally resulting in a higher cluster precision at the cost of cluster recall. A document-clustering technique with a PRT curve closer to the upper-right corner is more desirable.

Parameter Tuning

In the tuning experiments, we randomly chose the clustering results of ten subjects to determine appropriate values for parameters involved in each document-clustering technique investigated. The overall clustering effectiveness of each technique in the tuning experiments is calculated by averaging the cluster recall and cluster precision obtained from the ten subjects.

We first examine the effect of the number of features (*k*), ranging from 200 to 2000 in increments of 200, for document representation on the effectiveness of the content-based document-clustering technique. As we show in Figure 3 (to reduce the complexity of the figure, we show only a subset of values for *k*), the PRT curve of the content-based technique

moves toward to the upper-right corner as k increases. The content-based technique achieves the best clustering effectiveness when k is 2,000. Thus, we adopt 2,000 for k in subsequent experiments.

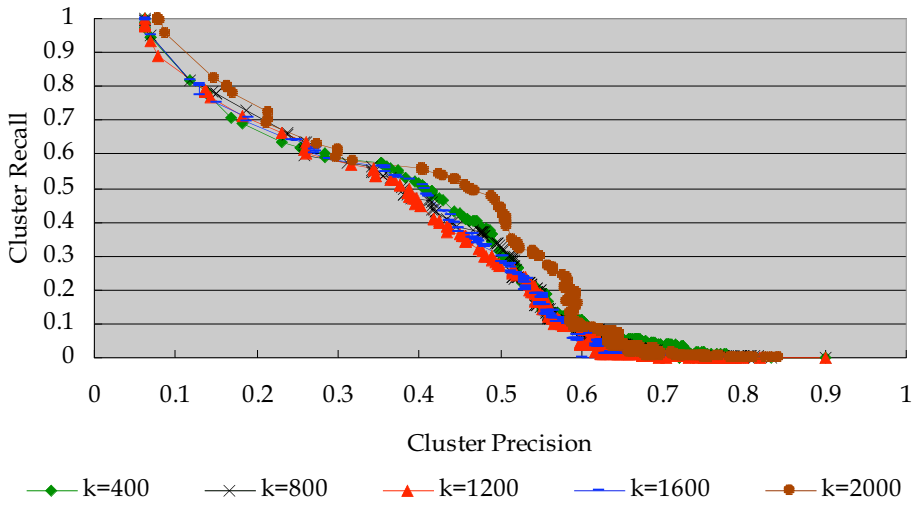

Figure 3: PRT Curves of the Content-based Document-Clustering Technique

Subsequently, we examine the effect of α (the threshold to determine whether a term should be expanded in the anchoring term expansion phase) for the CAC technique. We evaluate α ranging from 1 to 10 in increments of 0.5. As we show in the Figure 4 (only a subset of values for α is presented), the best clustering effectiveness attained by the CAC technique is when α equals to 2.5. Further increment or decrement of α degrades the performance of CAC technique. Thus, we adopt 2.5 for α in subsequent experiments.

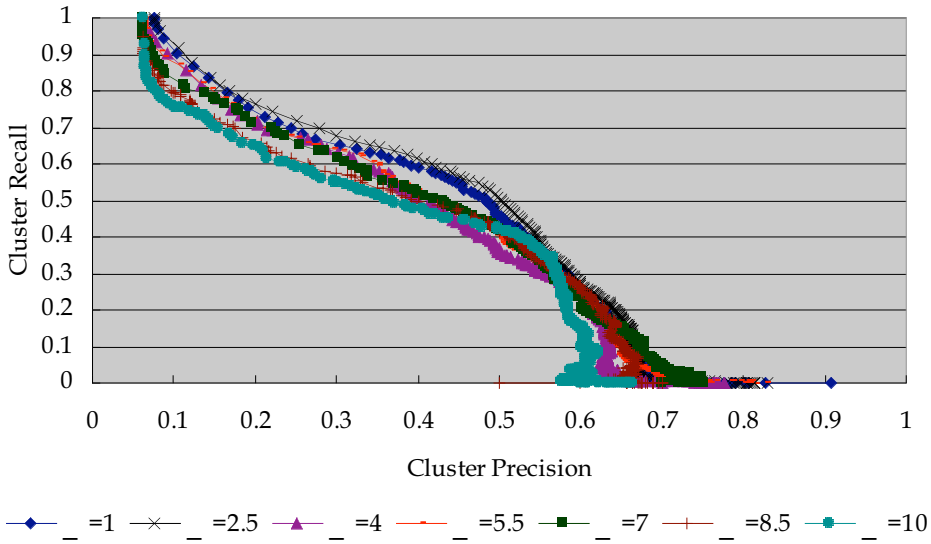

Figure 4: Effect of α for the CAC Technique

Comparative Evaluation

Using the parameter values determined previously, we evaluate the effectiveness of the

proposed CAC technique and its benchmark technique. In this experiment, all of the 33 subjects are used for evaluation purpose. The comparative evaluation result is shown in Figure 5. The proposed CAC technique achieves better clustering effectiveness than does the content-based document-clustering technique. This result suggests that our proposed CAC technique using the set of anchoring terms expanded by a statistical-based thesaurus constructed from a search engine for document representation can improve clustering effectiveness as measured by cluster recall and cluster precision. Because the PRT curve attained by each document-clustering technique forms a line in the cluster recall and cluster precision space, a statistical significant test between two lines is difficult, if not impossible. We therefore perform the significant test on the breakeven point attained by each technique. The breakeven point, an effectiveness measure commonly adopted by text categorization research (Sebastiani, 2002), is defined as the value at which cluster recall equals cluster precision. We first identify, for every subject, the breakeven point attained by each technique. The average breakeven point of the 33 subjects achieved by the CAC technique is 0.4962, noticeably higher than that attained by the content-based technique (i.e., 0.4679). We then conduct a paired *t*-test to test the statistical significance among the breakeven points of different document-clustering techniques. According to the test result, the proposed CAC technique significantly outperforms the content-based techniques at $p < 0.01$.

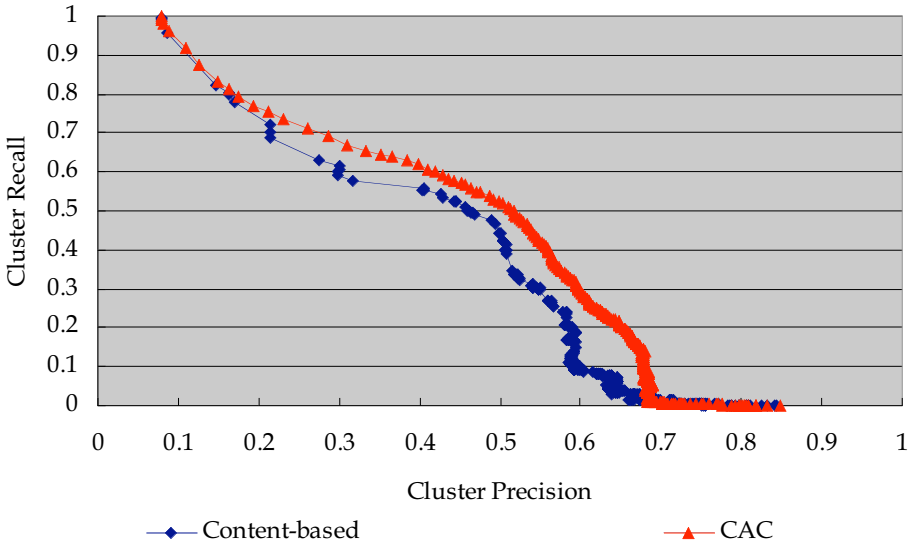

Figure 5: PRT Curves of Different Document-Clustering Techniques

In this study, we adopt a search engine (specifically, Google) for statistical-based thesaurus construction. A common alternative is to construct the statistical-based thesaurus from the target document corpus to be clustered. Thus, we further examine the difference on clustering effectiveness of the CAC technique with the use of the search-engine-based and the corpus-based statistical-based thesaurus, respectively. We employ the method proposed by Yang and Luk (2003) for the corpus-based statistical-based thesaurus construction. As Figure 6 illustrates, the search-engine-based method greatly outperforms the corpus-based method. A plausible explanation is that the statistical-based thesaurus constructed from a small-sized set of documents is limited in its vocabulary. Our analysis shows that, among the 33 sets of anchoring terms collected in this study, the average missing rate (i.e., the percentage of the set anchoring terms provided by a specific subject that are not present in the statistical-based thesaurus) of the corpus-based method is up to 32.15%, whereas the average missing rate of

the search-engine-based method is only 0.31%. As a result, more anchoring terms provided by a user do not appear in the corpus-based statistical-based thesaurus and cannot be expanded in the anchoring term expansion phase; thus, constraining the effectiveness of the CAC technique.

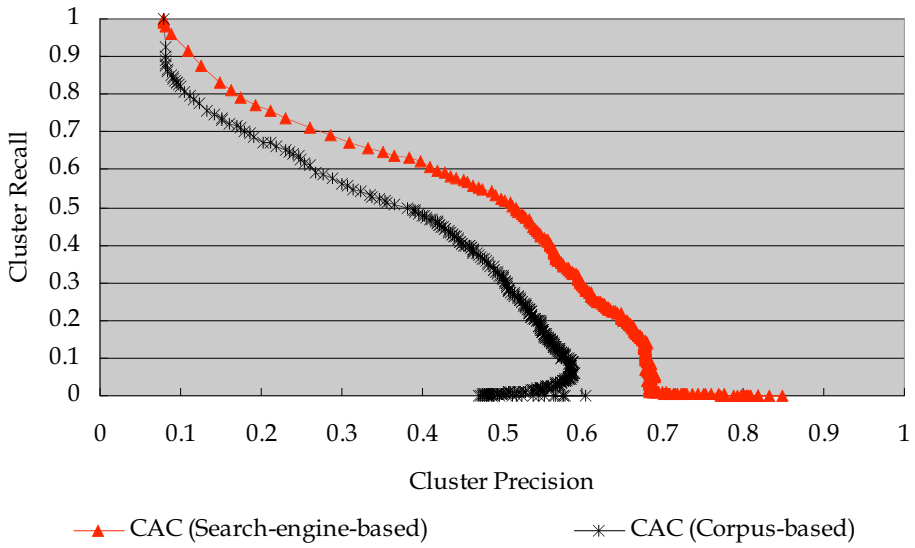

Figure 6: PRT Curves of Different Statistical-based Thesaurus Construction Methods

Conclusion and Future Research Directions

Existing document-clustering techniques typically generate a single set of clusters for all individuals without tailoring them to individuals' preferences and contexts and thus are unable to support context-aware document-clustering. Our research has been motivated by the importance of and need for context-aware document-clustering. In this study, we design and implement a Context-Aware document-Clustering (CAC) technique by taking into consideration a user's categorization preference relevant to the context of a target task and a statistical-based thesaurus constructed from the World Wide Web (WWW) for supporting context-aware document-clustering. Our empirical evaluation results suggest that our proposed CAC technique achieves better clustering results measured by cluster recall and precision than does the content-based document-clustering technique.

Some ongoing and future research directions are briefly discussed as follows. First, a user may have difficulty in giving a comprehensive set of anchoring terms. A future research direction is to evaluate and improve the performance of the CAC technique in the situation where only a partial set of anchoring terms are available. Second, this study only captures the statistical relevance between terms in the statistical-based thesaurus. However, there are still some other semantic relations between terms (e.g., synonymy, hyponymy, and hyperonymy), which may be beneficial for the anchoring term expansion task essential to the proposed CAC technique. Hence, it will be interesting and desirable to enhance our statistical-based thesaurus construction mechanism with a wider semantic coverage and to extend the CAC technique that exploits the diverse semantic relations for further improving clustering effectiveness. Last, our experimental study only includes research articles in our document corpus. Additional empirical evaluation involving documents from other domains (e.g., news, patents, etc.) is one of our future research directions.

Acknowledgments

This work was supported in part by National Science Council of the Republic of China under the grants NSC 95-2416-H-007-005 and NSC 95-2752-H-007-004-PAE.

References

Barreau, D.K., "Context as A Factor in Personal Information Management Systems," *Journal of the American Society for Information Science* (46:5), 1991, pp.327-339.

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

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

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

Case, D.O., "Conceptual Organization and Retrieval of Text by Historians: The Role of Memory and Metaphor," *Journal of the American Society for Information Science* (42:9), October 1991, pp.657-668.

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

Deogun, J. and Raghavan, V., "User-oriented Document Clustering: A Framework for Learning in Information Retrieval," In *Proceedings of the 9th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1986, pp.157-163.

Donovan, J., "Patrons Expectations about Collocation: Measuring the Difference between Psychologically Real and the Really Real," *Cataloging and Classification Quarterly* (13:2), 1991, pp.23-43.

El-Hamdouchi, A. and Willett, P., "Hierarchical document clustering using Ward's method." In *Proceedings of ACM Conference on Research and Development in Information Retrieval*, 1986, pp.149-156.

Jain, A.K., Murty, M.N., and Flynn, P.J., "Data Clustering: A Review," *ACM Computing Surveys* (31:3), 1999, pp.265-323.

Kim, H. and Lee, S., "A Semi-supervised Document Clustering Technique for Information Organization," In *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM)*, 2000, pp.30-37.

Kim, H. and Lee, S. "An Effective Document Clustering Method Using User-adaptable Distance Metrics," In *Proceedings of the 2002 ACM Symposium on Applied Computing*, 2002, pp.16-20.

Kwasnik, B.H., "The Importance of Factors that Are Not Document Attributes in the Organization of Personal Documents," *Journal of Documentation* (47), 1991, pp.389-398.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T., "Self-organizing Maps of Document Collections: A New Approach to Interactive Exploration," In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp.238-243.

Lakoff, G., *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, Chicago, IL, 1987.

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

Lin, C., Chen, H., and Nunamaker, J.F., "Verifying the Proximity and Size Hypothesis for Self-organizing Maps," *Journal of Management Information Systems* (16:3), Winter 1999-2000, pp.57-70.

Mackay, W.E., "Diversity in the Use of Electronic Mail: A Preliminary Inquiry," *ACM Transactions on Office Information Systems* (6:4), 1988, pp.380-397.

Mackay, W.E., "Responding to Cognitive Overload: Co-adaptation between Users and Technology," *Intellectica* (30:1), 2000, pp.177-193.

Pantel, P. and Lin, D., "Document Clustering with Committees," In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, pp.199-206.

Quillian, M.R., "Semantic Memory," In *Semantic Information Processing*, M. Minsky (ed.), The MIT Press, Cambridge, MA, 1968, pp.227-270.

Restorick, F.M., "Novel Filing Systems Applicable to An Automated Office: A State-of-the-Art Study," *Information Processing and Management* (22), 1986, pp.151-172.

Roussinov, D.G. and Chen, H., "Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques," *Decision Support Systems* (27:1-2), 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.

Sebastiani, F., "Machine Learning in Automated Text Categorization," *ACM Computing Surveys* (34:1), 2002, pp.1-47.

Talavera, L. and Bejar, J., "Integrating Declarative Knowledge in Hierarchical Clustering Tasks." In *Proceedings of the 3rd International Symposium on Intelligent Data Analysis*, 1999, pp.211-222.

Turney, P.D. and Littman, M.L., "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Transactions on Information Systems* (21:4), October 2003, pp.315-346.

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," In *Proceedings of Workshop on Very Large Corpora*, Columbus, Ohio, June 1993, pp.48-57.

Wei, C., Hu, P., 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.

Wei, C., Yang, C.S., Hsiao, H.W., and Cheng, T.H., "Combining Preference- and Content-based Approaches for Improving Document Clustering Effectiveness," *Information Processing and Management* (42:2), March 2006a, pp.350-372.

Wei, C., Chiang, R., and Wu, C., "Accommodating Individual Categorization Preferences: A Personalized Document Clustering Approach," *Journal of Management Information Systems* (23:2), Fall 2006b, pp.173-201.

Wei, C., Yang, C.S., and Hsiao, H.W., "A Collaborative-Filtering-Based Approach to Personalized Document Clustering," *Decision Support Systems*, forthcoming (2007).

Yang, Y. and Chute, C.G., "An Example-based Mapping Method for Text Categorization and Retrieval," *ACM Transactions on Information Systems* (12:3), 1994, pp.252-277.

Yang, C.C. and Luk J., "Automatic Generation of English/Chinese Thesaurus Based on a Parallel Corpus in Laws," *Journal of the American Society for Information Science and Technology* (54:7), 2003, pp.671-682.