December 2003

A Clustering-based Approach for Supporting Document-Category Integration

Chih-Ping Wei
National Sun Yat-Sen University

Tsang-Hsiang Cheng
National Sun Yat-Sen University

A Clustering-based Approach for Supporting Document-Category Integration

Chih-Ping Wei and Tsang-Hsiang Cheng

Department of Information Management
National Sun Yat-sen University, Kaohsiung, Taiwan, R.O.C.
cwei@mis.nsysu.edu.tw; cts@mail.stut.edu.tw

Abstract

Integration of relevant categorized documents into existent categories of an organization or individual is an important issue in the e-commerce era. Existing categorization-based approach for document-category integration (specifically, ENB) incurs several limitations, including homogeneous assumption on categorization schemes used by master and source catalogs and requirement for a large-sized master categories as training data. In this study, we developed a Clustering-based Category Integration (CCI) technique to address the problems inherent to categorization-based approach. Using the ENB as benchmarks, the empirical evaluation results showed that CCI appeared to improve the effectiveness of document-category integration accuracy in different integration scenarios and seemed to be less sensitive to the size of master categories than ENB.

Keywords

Category Integration, Catalog Integration, Document Clustering, Hierarchical Clustering, Naïve Bayes Classifier

1. Introduction

In the e-commerce era, organizations increasingly have participated in or shifted to the Internet environment for conducting online business transactions with suppliers and customers, gathering competitive intelligence information from various online sources, and sharing information and knowledge within or beyond organizational boundaries, etc. Such e-commerce applications generate and consume tremendous amount of online information that is typically available as textual documents. To manage the ever-increasing volume of online documents, organizations typically organize their documents into categories (or category hierarchies) to facilitate organizational document management and to assist subsequent information access and browsing by their users. The described use of categories is also commonly observed at the individual level, where individuals organize and archive into folders the documents, emails and favorite websites they currently hold.

Conceivably, an organization or individual constantly acquires relevant documents from various Internet sources. The categories used by an information provider for organizing its documents generally are dissimilar to or even different from those employed by the acquiring organization or individual. Consequently, integration of relevant categorized documents into existent categories deployed by the organization or individual has become a challenging issue for organizations and individuals.

Such document-category integration (hereafter referred to as category integration) need is pervasive in the emerging e-commerce environments since many websites are aggregators of

information from various online sources. For example, Yahoo! News (news.yahoo.com) aggregates news stories from multiple news providers (e.g., Reuters, Associated Press, Forbes.com, USA Today, etc.). Yahoo! News classifies business news into such categories as economy, stock markets, earning, personal finance, and industries commentary, while Forbes.com, for instance, includes for business news such categories as manufacturing, technology, commerce, services, energy, and health care. Facing non-identical categories used by various news providers, Yahoo! News evidently requires a category integration mechanism for automatically and effectively integrating categorized news stories from news providers into its categories. The described category integration is also essential for integrating product categories along supply chains. For example, a distributor maintains an online product catalog to be accessed and ordered by its customers (retailers or consumers). Imaginably, a large distributor typically has hundreds, if not thousands, of suppliers each of which has its own product categorization scheme. As a result, the distributor must integrate individual product catalogs from its suppliers (Stonebraker & Hellerstein 2001; Agrawal & Srikant 2001).

Previous research formulated category integration as a text categorization problem. With such formulation, a set of document categories is designated as the master catalog *M*, while another set of categories assumes the role of source catalog *S*. Accordingly, the category integration problem is to assign each document in *S* into the most appropriate category in *M* or into a new category if the document cannot be properly assigned to any predefined category in *M* (Agrawal & Srikant 2001). Using a text categorization technique, the categorization-based approach for category integration involves the learning of a classification model based on the categorized documents in *M* as the training instances. Subsequently, using the classification model induced, each document in *S* is assigned into the most appropriate category or the new category in *M*. This straightforward approach, however, completely ignores the categorization information present in *S*. Agrawal and Srikant (2001) argued that the accuracy of category integration could be improved by taking into consideration the implicit categorization information of *S*. Their intuition is that if two documents belong to the same category in *S*, these two documents are more likely to belong to the same category in *M*. Accordingly, they developed an Enhanced Naï ve Bayes (ENB) classifier for improving the accuracy of category integration.

Although their empirical results showed that ENB could improve the accuracy of category integration, the categorization-based approach incurs several limitations. First, to be effective, the categorization-based approach for category integration assumes that the categorization schemes used by both catalogs (i.e., master and source) are homogeneous (Agrawal & Srikant 2001). However, if the categorization schemes used by the two catalogs are completely orthogonal to each other, the implicit categorization information of the source catalog will not help the categorization-based approach to improve integration accuracy. Secondly, by its inductive nature, the categorization-based approach for category integration typically requires a fairly large training data set in order to achieve satisfactory integration accuracy. This requirement may be reasonable for large-sized organizations where the number of documents in each master category may be large. However, such data size requirement may not be realistic for small-sized organizations or individuals. In this case, ENB may not be effective for integrating documents in the source categories into appropriate categories in the master catalog.

To address the above-mentioned limitations of the categorization-based approach, this study attempts to propose a clustering-based approach for category integration. In essence, document clustering groups documents into clusters, each of which contains documents similar in their content (Voorhees 1986; El-Hamdouchi & Willett 1986). Thus, by dividing each source category into several subcategories, each of which contains documents of a more

specific scope and increased cohesion, the clustering-based approach may be capable of handling heterogeneous categorization schemes used by master and source catalogs. Subsequently, using a document clustering technique, the two catalogs are integrated by merging source subcategories into most similar master categories. Evidently, as an unsupervised learning method, the clustering-based approach for category integration would be less sensitive to the sizes of master categories than the categorization-based approach.

Specifically, in this study, we designed and implemented a Clustering-based Category Integration (CCI) technique. Experimentally, we evaluated CCI's effectiveness using that achieved by the categorization-based approach (i.e., ENB) as a benchmark. The remainder of the paper is organized as follows. Section 2 reviews relevant previous research on category integration and document clustering techniques. Section 3 details the proposed technique (i.e., CCI). An empirical evaluation using synthetic datasets will be conducted and important experimental results will be discussed in Section 4. The paper is concluded with a summary and some future research directions in Section 5.

2. Literature Review

As mentioned, the categorization-based approach employs and enhances the Naive Bayes classification algorithm for category integration. Hence, in the following, a brief review on the Naive Bayes and the Enhanced Naive Bayes (ENB) algorithms will be provided. Furthermore, since the proposed CCI technique takes the document clustering approach, the general process and existing techniques of document clustering will be highlighted.

2.1 Naive Bayes Classification for Category Integration

The Naive Bayes classifier uses the joint probabilities of words and categories to estimate the probabilities of categories given a document. It is based on the assumption of word independence (i.e., the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category). The Naive-Bayes classifier estimates the posterior probability of the category C_i given a document d (i.e., $p(C_i|d)$) via Bayes rule as (Mitchell 1996):

$$p(C_i|d) = \frac{p(d|C_i) \times p(C_i)}{p(d)}$$

where $p(C_i)$ is the probability that C_i occurs, $p(d)$ is the probability that d occurs, and $p(d|C_i)$ is the conditional probability that d occurs given C_i.

$p(C_i)$ can be estimated by the number of documents in C_i divided by the total number of documents in the dataset. On the other hand, $p(d)$ can be ignored when estimating $p(C_i|d)$ since it is identical for all categories and the relative probability of the categories can be used to determine d's category assignment. Due to the word independence assumption, $p(d|C_i)$ can be derived as $p(d|C_i) = \prod_{f \in d} p(f|C_i)$, where f represents a word in d. The maximum likelihood estimate is typically employed for estimating $p(f|C_i)$ as $freq(C_i, f)/freq(C_i)$ where $freq(C_i, f)$ is the number of occurrences of f in all documents in C_i, and $freq(C_i) = \sum_{f \in C_i} freq(C_i, f)$ refers to the total number of words (counting multiple occurrences) in C_i.

Evidently, if a word f in a document d does not appear in any documents in C_i, $p(f|C_i)$ and, thus, $p(d|C_i)$ becomes 0. Imaginably, d is likely to contain some words that are not present in any category, leading to $p(d|C_i)$ equal to 0 for all categories C_i. In this case, due to the existence of such words in d, d cannot be assigned to any category. To avoid such situation caused by the

undesired property of the described estimate for $p(f|C_i)$, Lidstone's law of succession was proposed to smooth the maximum likelihood estimate (Agrawal et al. 2000). For $\lambda \geq 0$, $p(f|C_i)$ is estimated as:

$$p(f|C_i) = \frac{freq(C_i, f)+\lambda}{freq(C_i)+\lambda|V|}$$

where $|V|$ is the number of distinct words in all documents across categories.

To improve accuracy of category integration, Agrawal and Srikant (2001) extended the described Bayes rule based on the rationale that if two documents belong to the same source category, they are more likely to belong to the same master category. Accordingly, to determine the category assignment for a document d in S, the Enhanced Naï ve Bayes (ENB) classification algorithm employs the posterior probability of the category C_i in M given a document d belonging to the category S_j in S as:

$$p(C_i|d, S_j) = \frac{p(d|C_i)\times p(C_i|S_j)}{p(d|S_j)}$$

$p(d|S_j)$ can be ignored when estimating $p(C_i|d, S_j)$ since $p(d|S_j)$ is the same for all C_i. To estimate $p(C_i|S_j)$, the documents in S_j are first classified using the standard Naï ve Bayes classifier. Subsequently, using the majority principle, $p(C_i|S_j)$ will be increased if the majority of documents in S_j is predicted to be in C_i. Correspondingly, the estimate for $p(C_i|S_j)$ is revised as:

$$p(C_i|S_j) = \frac{|C_i|\times(\text{number of documents in } S_j \text{ predicted to be in } C_i)^w}{\sum_{k=1}^{n} (|C_k|\times(\text{number of documents in } S_j \text{ predicted to be in } C_k)^w)}$$

where $w \geq 0$ and $|C_i|$ is the number of documents in the master category C_i.

2.2 Document Clustering

Document clustering is to automatically organize a collection of documents into distinct groups of similar documents and to discern general themes hidden in the corpus. The general purpose of document clustering broadly consists of several steps, including feature extraction, feature selection, document representation, and clustering. The feature extraction and selection steps extract a set of features from the target documents and select most representative features that will be used for representing the target documents in the document representation step. Typically, feature extraction commences with the parsing of each target document to produce a list of nouns or noun phrases commonly referred to as features. Afterward, a set of specified stop words that are non-semantic bearing words will be excluded from the feature set. Following extraction is feature selection, which reduces the size of the extracted feature set. Feature selection has important implications for the subsequent clustering efficiency and effectiveness (Dumais, Platt, Heckerman & Sahami 1998; Roussinov & Chen 1999). According to the top-k selection method, the k features with the highest selection metric scores are selected. Previous research commonly employed such feature selection metrics as TF (which denotes the occurrence frequency of a particular term in the document collection), TF×IDF (where IDF denotes the inverse document frequency), and their hybrids.

In the document representation step, each target document is represented using the feature set selected from the previous step. That is, each target document is described by a vector space jointly defined by the k features selected previously. A review of prior research suggests the prevalence of several representation methods that include binary (which indicates the presence or absence of a feature in a document), TF, and TF×IDF (Larsen & Aone 1999; Roussinov & Chen 1999).

In the clustering step, documents are grouped into clusters, based on the selected features and their respective values for each target document. Common approaches for document clustering include partitioning-based (e.g., *k*-means), hierarchical (e.g., hierarchical agglomerative clustering and hierarchical divisive clustering algorithms), and Kohonen neural network (El-Hamdouchi & Willett 1986; Lagus et al. 1996; Larsen & Aone 1999; Roussinov & Chen 1999; Voorhees 1986)

3. Clustering-based Category Integration Technique

The Clustering-based Category Integration (CCI) technique takes as inputs two catalogs (where one catalog is designated as the master catalog and the other is the source catalog) together with respective categorized documents, and then determines or produces integrated categories. For purposes of the intended feasibility assessment, the current design of CCI concentrates on integrating catalogs with a flat structure. That is, the categories in each catalog, master and source, are organized in a flat set rather than in a hierarchical structure. As depicted in Figure 1, the overall process of CCI consists of extraction of categorization scheme, source category decomposition, and category merging phases.

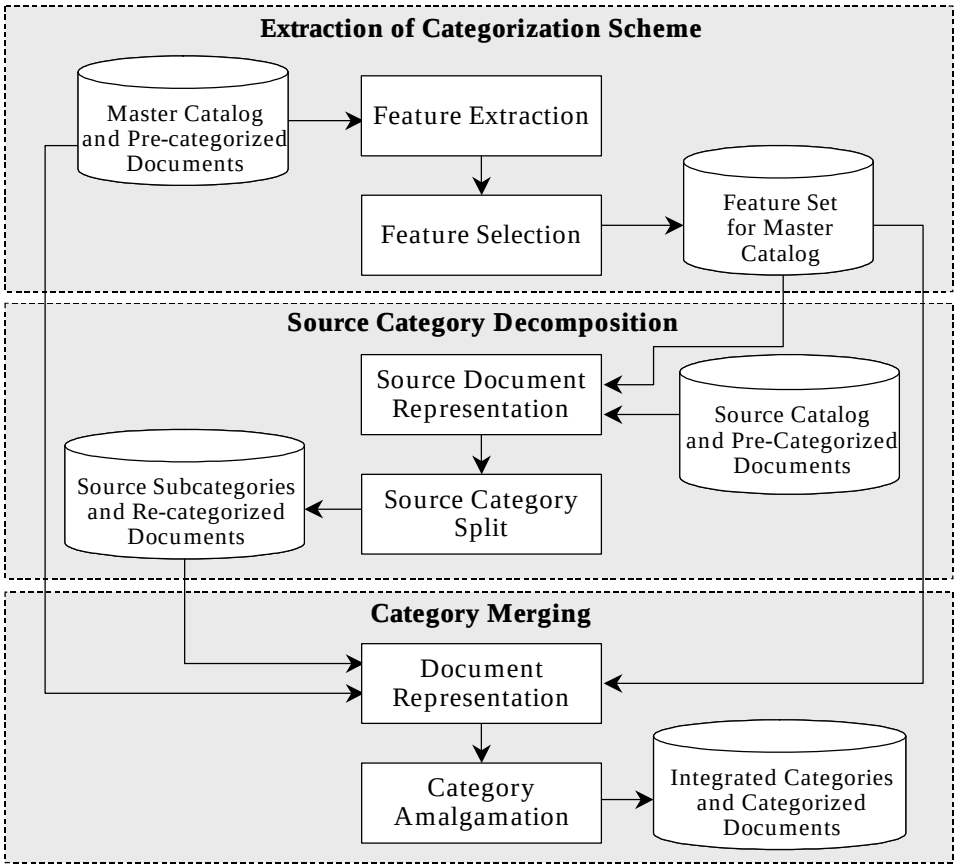

Figure 1: Process of CCI Technique

3.1 Extraction of Categorization Scheme

This phase is to select a set of features that appropriately differentiate categories in the master catalog. Using an adequate feature selection method, the set of features extracted would capture and reflect the categorization scheme of the master catalog.

Feature extraction: CCI first extracts a set of features (including nouns and noun phrases) from the (categorized) documents in the major catalog *M*. We adopted the rule-based part-of-speech tagger developed by Brill (1992; 1994) to syntactically tag each word in the master documents. Subsequently, we employed the approach proposed by Voutilainen (1993) to implement a noun-phrase parser for extracting noun phrases from each syntactically tagged document. As result, each master document is now represented as a set of nouns and noun phrases.

Feature selection: Subsequently, the weighted average c^2 statistic method (Yang & Pedersen 1997) was adopted for feature selection. The c^2 statistic, measuring the dependence between a feature *f* and a category C_i, inclines to 0 when *f* and C_i are independent. Let N_{r+} be the number of documents in C_i where *f* occurs, N_{r-} be the number of documents in C_i where *f* does not appear, N_{n+} is the number of documents in the categories other than C_i where *f* occurs, N_{n-} is the number of documents in the categories other than C_i where *f* does not appear, and *N* is the total number of documents in the master categories. The c^2 statistic of *f* relevant to C_i is defined as:

$$c^2(f, C_i) = \frac{N \times (N_{r+}N_{n-} - N_{r-}N_{n+})^2}{(N_{r+}+N_{r-})(N_{n+}+N_{n-})(N_{r+}+N_{n+})(N_{r-}+N_{n-})}$$

Once the c^2 statistic of *f* relevant to each category C_i in the master catalog *M* is derived, the overall c^2 statistic of *f* to all categories in *M* is estimated using the weighted average scheme. That is, $c^2(f, M) = \sum_{C_i \in M} p(C_i) \times c^2(f, C_i)$ where $p(C_i)$ is estimated by the number of documents in C_i divided by *N*. Accordingly, the *k* features with the highest weighted c^2 statistic scores are selected as the categorization scheme of *M*.

3.2 Source Category Decomposition

This phase divides each source category into several more cohesive subcategories in order to reduce the heterogeneity between the categorization schemes of the master and source catalogs. Two tasks are performed in this phase, including source document representation and source category split.

Source Document Representation: In order to produce source subcategories whose categorization scheme is comparable to that of the master catalog, the categorization scheme of the master catalog extracted previously is employed for representing each source document. In this study, the binary or term frequency (TF) representation method was employed. That is, each source document d_i is described by a feature vector $\vec{d}_i = <r_{i1}, r_{i2}, \ldots, r_{ik}>$, where r_{ij} is 1 if the feature f_j appears in d_i, and 0 otherwise. On the other hand, r_{ij} in the TF representation method is the frequency of f_j in d_i.

Source Category Split: We employed the hierarchical divisive clustering (HDC) method for decomposing each source category into a set of subcategories. Choice of the HDC method over other partitioning-based techniques was made primarily because of its advantage in that the number of clusters need not be pre-specified and that the number of clusters can be increased (or decreased) by simply moving down (or up) the resultant clustering hierarchy. Specifically, for each source category, the HDC algorithm (Kaufman & & Rousseeuw 1990) starts with all documents in one cluster, and then subdivides the category into two smaller clusters until intra-cluster similarity of each cluster is above a pre-defined similarity threshold (a_D). 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 d_i and $|\vec{d_i}|$ is the length of $\vec{d_i}$.

3.3 Category Merging

The category merging phase performs the merging of the source subcategories into the master categories and generates a set of integrated categories and respective documents in each integrated category. This phase involves two tasks, including document representation and category amalgamation.

Document Representation: All of the documents, master and source, are represented using the feature set extracted from and selected for the master catalog. As with the source category decomposition phase, the binary or TF method was adopted for document representation.

Category Amalgamation: This task is to merge the source subcategories into the master categories. The criterion for such merging decision is based on the similarity between source subcategories and master categories. Besides, additional merging constraints are needed.

Rule 1: Directly merging two source subcategories is not allowed since it is not consistent with the goal of category integration. However, two source subcategories can finally be joined together if both of them are merging into the same master category.

Rule 2: Merging of two categories each of which encompasses a master category is prohibited since the master catalog is assumed to be unchanged in this study. However, if this constraint is too strict for some category integration applications, it can easily be modified or relaxed to accommodate other integration requirements.

Due to these category merging constraints, use and extension of the hierarchical agglomerative clustering (HAC) algorithm seems to be more appropriate and viable than other clustering algorithms. The extended HAC algorithm starts with as many clusters as there are (sub)categories in the master and source catalogs. Subsequently, two most similar clusters whose merging does not violate any of the constraints specified previously will be merged to form a new cluster. In this study, the group-average link method (i.e., the average similarity between all inter-cluster pairs of documents) was employed to measure the similarity between two clusters. This merging process continues until the similarity of a permissible cluster-merger is less than a pre-defined similarity threshold (a_I). For those source subcategories that cannot be merged into any master categories, these standalone source subcategories will be deposited into a new category (called the residual category) in the master catalog. Generation of appropriate categories for those documents in the residual category is an essential but challenging research issue that will be addressed in our future work.

Upon the completion of merging the source subcategories into the master categories via the extended HAC algorithm, a set of integrated categories and their respective documents will be generated.

4. Empirical Evaluations

This section reports the empirical evaluation of the proposed CCI technique, using the categorization-based category integration approach (specifically, ENB) as performance benchmarks.

4.1 Data Sets

The document collection for evaluation purpose was collected from CiteSeer Scientific Literature Digital Library website (http://citeseer.nj.nec.com/) and consisted of 384 research articles related to information systems and technologies. For each article, only the title, abstract and keywords were used in this evaluation study. The 384 research articles were manually classified into nine categories including data compression (28 documents), data mining (62), electronic commerce (32), information retrieval (45), network protocol (18), robotics (28), security (60), wireless network (61), and XML (50).

The nine document categories were considered as "true" categories and served as the foundation for synthetically generating two catalogs. Accordingly, three scenarios where category integration were needed were simulated in this study:

1. Homogeneous (*M9S9*) scenario: In this scenario, the categorization schemes of the master and source catalogs are intended to be homogeneous. Specifically, a certain percentage of documents (e.g., 50%) in the "true" categories were selected randomly and assigned into the master catalog whose category structure was the same as the "true" categories, while the remaining documents were placed in the source catalog. To preserve the categorization scheme of the master catalog in the source catalog, the category structure of the source catalog also imitated that of the "true" categories. As a result, two 9-category catalogs with homogeneous categorization schemes were created.
2. Comparable (*M9S4*) scenario: The master and source categories were initially created by the same manner as described in the homogeneous scenario. To simulate comparable categorization schemes between two catalogs, the initial nine source categories were randomly combined into four source categories. As such, a 9-category master catalog and a 4-category source catalog were created.
3. Heterogeneous (*M9S10*) scenario: The master and source categories were again initially created by the same manner as described in the homogeneous scenario. Subsequently, the initial category structure of the source catalog was discarded and their documents were randomly and evenly assigned into ten categories. As such, two catalogs (i.e., one with 9 master categories and another with 10 source categories) were created.

4.2 Evaluation Procedure and Criteria

When a synthetic dataset was randomly generated for each of the described scenarios, CCI or ENB was applied and its resulting integrated categories were then compared with the "true" categories. To minimize the potential biases resulting from the randomization process, the described synthetic dataset creation process was performed thirty times and the overall performance for each category integration technique investigated was estimated by averaging the performance estimates obtained from the 30 individual trials.

Since the three scenarios simulate the circumstances under which all documents in the source categories can be integrated into the master categories, the effectiveness of a category integration technique under discussion thus, can be measured by integration accuracy rate (*IA*), defined as:

$$IA = \frac{m}{|S|}$$

where m is the number of documents in the source catalog S that are correctly classified and $|S|$ is the number of documents in S.

4.3 Parameter Tuning for ENB

As depicted in Section 2.1, ENB involves two parameters, λ and *w*. We used the comparable scenario and a 50/50 document-distribution between master and source catalogs for generating tuning datasets. When tuning λ, we set *w* to zero (where ENB is degenerated to the standard Naï ve Bayes classifier). We examined different values for λ, ranging from 0.1 to 1.5. Based on empirical results, we decided on 0.6 for λ, which appeared to attain the highest integration accuracy, for subsequent experiments.

Using 0.6 for λ, we then determined an appropriate value for *w* of ENB. For possible values for *w*, an exponentially increasing series (e.g., 0, 1, 3, 10, 30, 100, 300, 1000) suggested by Agrawal and Srikant (2001) was investigated. As shown in Figure 2, under the homogeneous scenario, an increase of *w* resulted in an improvement on integration accuracy. Such improvement eventually flattened out when *w* reached 100. For the comparable scenario, integration accuracy of ENB increased and then decreased as *w* increased. In this scenario, the highest integration accuracy attained by ENB appeared when *w* was 10. Finally, in the heterogeneous scenario, as *w* increased, the resulting integration accuracy decreased progressively. This result suggested that setting 0 for *w* would be appropriate. Hence, *w* of 100, 10 and 0 were selected for the homogeneous, comparable, and heterogeneous scenarios, respectively.

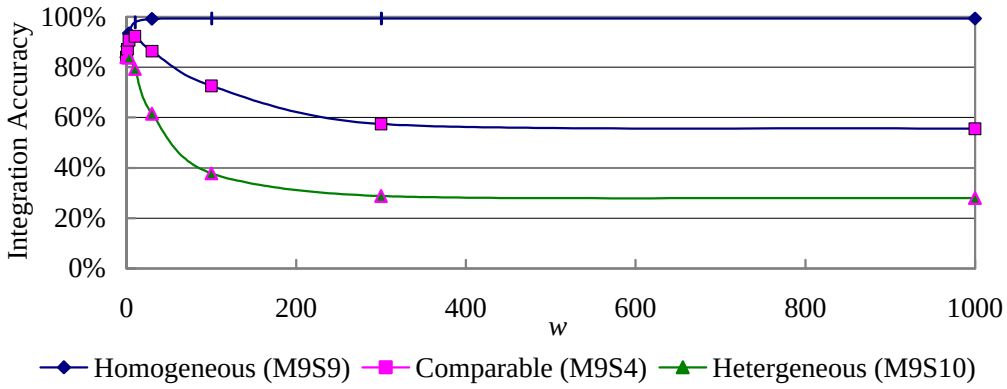

Figure 2: Tuning Results for w of ENB

4.4 Parameter Tuning Experiments for CCI

The parameters or design choices involved in CCI include the number of features (*k*), document representation method, the similarity threshold (a_D) for source category split, and the similarity threshold (a_I) for category amalgamation. Since all documents in the source categories can be integrated into the master categories in any of the scenarios designed, setting a_I as 0 would be appropriate and desirable. We used a 50/50 document-distribution for generating tuning datasets for each scenario. Our empirical results showed that, across different scenarios, the binary representation would achieve higher integration accuracy of CCI than the TF representation method. In addition, we examined different number of features (*k*), ranging from 50 to 500 at increments of 50. Empirical results suggested that setting *k* as 250 would appropriate for the three scenarios investigated. Thus, for subsequent experiments, we selected *k* of 250 and the binary representation method for document representation.

We examined different values for a_D, ranging from 0 to 1 at increments of 0.1. Apparently, an increase of a_D favors a finer decomposition for each source category. As shown in Figure 3,

when the categorization schemes were homogeneous, no source decomposition (i.e., 0 for a_D) led to a perfect integration result. However, in the comparable scenario, when a_D increased from 0 to 0.3, integration accuracy of CCI improved. However, when a_D was greater than 0.3, integration accuracy of CCI degraded. Finally, in the heterogeneous scenario, CCI's integration accuracy improved as a_D increased from 0 to 0.6. When a_D was higher than 0.6, integration accuracy of CCI generally became stable. Overall, a_D as 0, 0.3, and 0.6 appeared to be most appropriate for the *M9S9*, *M9S4*, and *M9S10* scenarios, respectively.

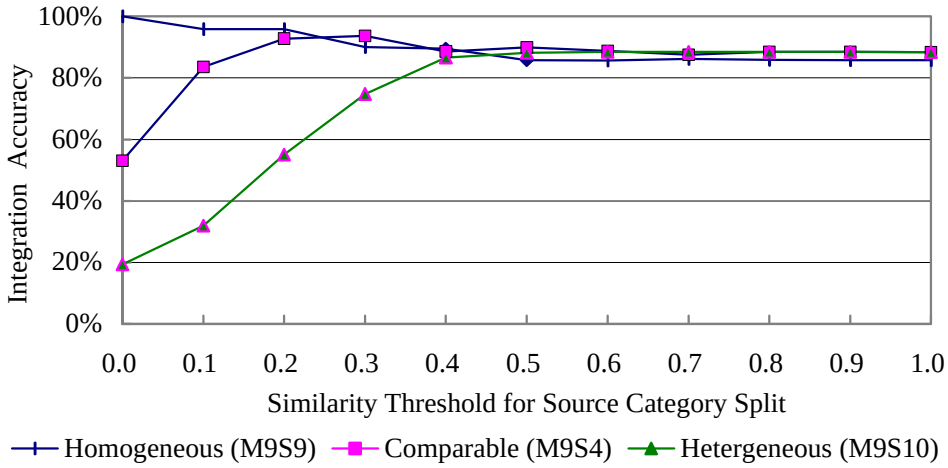

Figure 3: Tuning Results for α_D of CCI

4.5 Comparative Evaluation

Using the experimentally determined parameter values, a comparison of integration accuracy of CCI and ENB under different scenarios is shown in Table 1. As shown, in all scenarios, integration accuracy achieved by CCI was higher than that attained by ENB. At a significance level of 95%, the difference was statically significant. Results from this comparative evaluation suggested that CCI was more effective than ENB.

	Scenario		
	Homogeneous	Comparable	Heterogeneous
	Average Integration Accuracy		
CCI	100%	93.69%	88.48%
ENB	99.37%	92.39%	84.04%
	t-value		
CCI-ENB	3.515***	3.575***	28.450***

: significant at the 0.05 level and *: significant at the 0.01 level

Table 1: Comparative Evaluation Results

4.6 Effects of Data Sizes

This evaluation task was designed to evaluate how well CCI performed at different sizes of master catalogs, using those achieved by ENB as benchmarks. Specifically, when generating a synthetic dataset, the documents in the "true" categories were distributed to the master and

source catalogs using different distributions. We varied the percentages of documents allocated to a master catalog from 50% to 10% at decrements of 10%. Using the experimentally determined parameter values, integration accuracy of CCI and ENB under different document-distribution designs and scenarios was estimated respectively.

As shown in Table 2, a decrease of size of master catalog generally has negative effects on integration accuracy of both CCI and ENB; however, ENB appeared to be more sensitive than CCI. In the homogeneous scenario, when the document-distribution for master catalogs decreased from 50% to 10%, integration accuracy achieved by CCI dropped by 8.54% (i.e., from 100% to 91.46%). However, over the same range of document-distribution in the homogeneous scenario, integration accuracy of ENB decreased by 30.96% (i.e., from 99.37% to 68.41%). When the scenario increased its heterogeneity, the performance of ENB declined more rapidly than that of CCI as the document-distribution for master catalogs decreased. The performance reduction of ENB was 42.97% and 31.02%, while that of CCI was 6.38% and 3.92% in the comparable and heterogeneous scenarios respectively. Furthermore, in most of the document-distribution and scenario combinations, CCI appeared to outperform ENB in integration accuracy. Such empirical results suggested that CCI would be less susceptible to the sizes of master categories than ENB.

Document-Distribution (Master-Source)	Technique	Scenario		
		Homogeneous	Comparable	Heterogeneous
50-50	CCI	100%	93.69%	88.48%
	ENB	99.37%	92.39%	84.04%
40-60	CCI	99.47%	89.47%	84.56%
	ENB	98.98%	90.32%	84.17%
30-70	CCI	94.04%	88.96%	85.23%
	ENB	97.60%	89.51%	85.10%
20-80	CCI	91.46%	87.31%	84.89%
	ENB	96.69%	81.94%	73.50%
10-90	CCI	95.36%	91.90%	86.88%
	ENB	68.41%	49.42%	54.08%

Table 2: Data Size Sensitivity

5. Conclusions

Integration of relevant categorized documents into existent categories deployed by an organization or individual is an important issue in the e-commerce era. Existing categorization-based approach for category integration (specifically, ENB) incurs several limitations, including homogeneous assumption on categorization schemes used by master and source catalogs and requirements for large-sized master categories as training data. In this study, we developed a Clustering-based Category Integration (CCI) technique to address the problems inherent to the categorization-based approach. Using ENB as benchmarks, the empirical evaluation results showed that the proposed CCI technique appeared to improve the effectiveness of category integration accuracy in various integration scenarios and seemed to be less sensitive to the sizes of master categories than the categorization-based approach.

Some ongoing and future research directions are briefly discussed as follows. This research

concentrated on document categories organized non-hierarchically. However, it is common that users organize their folders in a hierarchical structure. Hence, the proposed CCI technique has to be extended for hierarchical category structures. On the other hand, the current design of CCI was mainly based on the content analysis. Incorporation of semantic analysis or external knowledge (e.g., users' browsing behaviors) into CCI would be essential and have a profound impact on category integration research. Finally, this study did not address the generation of appropriate categories for those source documents that cannot properly be integrated into any master categories. New category generation is an essential but challenging research issue since the categorization scheme used for deriving new categories need to be comparable, if not homogeneous, to that used by the existing master catalog.

Acknowledgment

This work was supported in part by National Science Council of the Republic of China under the grant NSC 90-2416-H-110-022 and MOE Program for Promoting Academic Excellence of Universities under the grant 91-H-FA08-1-4.

References

Agrawal, R, Bayardo, R, & Srikant, R (2000), "Athena: Mining-based Interactive Management of Text Databases," *Proceedings of the Seventh Conference on Extending Database Technology (EDBT00)*, pp.365-379.

Agrawal, R & Srikant, R (2001), "On Integrating Catalogs," *Proceedings of the tenth international conference on World Wide Web*, pp.603-612

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

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

Dumais, S, Platt, J, Heckerman, D & Sahami, M (1998), "Inductive Learning Algorithms and Representations for Text Categorization," *Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management (CIKM '98)* , pp.148-155.

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

Kaufman, L & Rousseeuw, PJ (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Inc.

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

Lagus, K, Honkela, T, Kaski, S, & Kohonen, T (1996), "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.

Mitchell, T (1996), *Machine Learning*, McGraw Hill.

Roussinov, D & Chen, H (1999), "Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques," *Decision Support Systems*, Vol. 27, No. 1-2, pp.67-79.

Stonebraker, M & Hellerstein, JM (2001), "Content Integration for E-Business," *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp.552-560.

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

Voutilainen, A (1993), "Nptool: A Detector of English Noun Phrases," *Proceedings of Workshop on Very Large Corpora*, Ohio.

Yang, Y & Pedersen, JO (1997), "A Comparative Study on Feature Selection in Text Categorization," *Proceedings of 14th International Conference on Machine Learning*, pp.412-420.