Personalized Document Clustering with Dual Supervision

Yeming Hu
Dalhousie University
Faculty of Computer Science
6050 University Avenue
Halifax, Canada
yeming@cs.dal.ca

Evangelos E. Milios
Dalhousie University
Faculty of Computer Science
6050 University Avenue
Halifax, Canada
eem@cs.dal.ca

James Blustein
Dalhousie University
Faculty of Computer Science
and School of Information
Management
Halifax, Canada
jamie@cs.dal.ca

Shali Liu
Dalhousie University
Faculty of Computer Science
6050 University Avenue
Halifax, Canada
shali@cs.dal.ca

ABSTRACT

The potential for semi-supervised techniques to produce personalized clusters has not been explored. This is due to the fact that semi-supervised clustering algorithms used to be evaluated using oracles based on underlying class labels. Although using oracles allows clustering algorithms to be evaluated quickly and without labor intensive labeling, it has the key disadvantage that oracles always give the same answer for an assignment of a document or a feature. However, different human users might give different assignments of the same document and/or feature because of different but equally valid points of view. In this paper, we conduct a user study in which we ask participants (users) to group the same document collection into clusters according to their own understanding, which are then used to evaluate semi-supervised clustering algorithms for user personalization. Through our user study, we observe that different users have their own personalized organizations of the same collection and a user's organization changes over time. Therefore, we propose that document clustering algorithms should be able to incorporate user input and produce personalized clusters based on the user input. We also confirm that semi-supervised algorithms with noisy user input can still produce better organizations matching user's expectation (personalization) than traditional unsupervised ones. Finally, we demonstrate that labeling keywords for clusters at the same time as labeling documents can improve clustering performance further compared to labeling only documents with respect to user personalization.

DocEng'12, September 4–7, 2012, Paris, France.
Copyright 2012 ACM 978-1-4503-1116-8/12/09 ...$15.00.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; H.1.2 [**User/Machine Systems**]: [Human Factors]; I.7.1 [**Computing Methodologies**]: Document and Text Processing—*Document Management*

General Terms

Algorithms, Human Factors, Design, Management

Keywords

User Supervision, Document Supervision, Feature Supervision, User Interface, Personalization

1. INTRODUCTION

Nowadays, academic researchers maintain a personal library of papers related to their research and courses, downloaded from digital libraries such as Association for Computing Machinery (ACM) digital library[1]. While those papers might be placed into different categories (folders) when they were downloaded, the categories are generally quite coarse. Even worse, papers with different topics might be put into the same folder only for temporary convenience. In fact, even if users categorize the papers appropriately at one time, they might change their mind later on and want to organize the papers in another manner. In addition, researchers might like one organization for their research but another one for preparing their courseware. Therefore, the organization of the personal library should be easily changed over time based on user's needs.

Clustering techniques are often employed to group a document collection into different topics. Unsupervised clustering does not require any user effort. However, the users may not be satisfied with the universal output since it does not reflect the individual user's point of view and completely ignores personalization. Semi-supervised clustering incorporates prior information, e.g., user input, into clustering algorithms and normally can produce better quality of clusters.

[1] `http://dl.acm.org/`

User input is generally provided through user supervision. With respect to document clustering, there are two types of supervision, i.e., document supervision and feature supervision. In document supervision, users provide document-level user input such as labeling a few representative documents for each cluster [2] or identifying relationship between two documents, i.e., "must-link" and "cannot-link" [21]. In feature supervision, users provide feature-level user input such as assigning a few keywords for each cluster [12] or identifying the features (words) which are useful for clustering [8, 9]. The semi-supervised clustering algorithms can also produce personalized clusters if combined with user inputs from individual users.

The previous semi-supervised clustering algorithms were all experimentally evaluated using oracles. Oracles are based on the underlying class labels of standard datasets. In the case of document supervision, two documents are put into the same cluster or identified as "must-link" by the oracle if they have the same class labels. Otherwise, they are identified as "cannot-link" and must end up in different clusters. With respect to feature supervision, a feature oracle is constructed using feature selection techniques such as χ^2 or information gain based on the underlying labels of documents. The constructed feature oracle determines whether a feature is useful for clustering and which cluster the feature should be assigned to. By using oracles, a new semi-supervised clustering algorithm can be evaluated and verified easily and quickly. However, there are two main disadvantages using oracles to evaluate semi-supervised algorithms. First, oracles always give the correct assignments of documents into clusters or "must-link" and "cannot-link". In real situation, human users can easily make mistakes in assigning documents. Therefore, the semi-supervised algorithms should be tested under noisy supervision, e.g., two documents are placed into the same cluster when they are not meant to. The same problem exists with feature supervision that a user can pick a useless feature or even assign one cluster's feature to another one especially when there are overlaps between clusters. Although one might claim that noise can be injected into oracle decisions [9], the probability method used to create the feature oracle may not be able to simulate a user's complicated decision process. Second, oracles constructed for one dataset always assign the same label for the same document or the same feature. Assume we have two papers and one talks about programming languages and the other is about software debugging. A document oracle based on underlying class labels will always give the same assignments on whether those two papers should be placed into the same cluster. However, one human user can assign them into the same cluster "software engineering" while another one would like to put them into two clusters, i.e., "languages" and "debugging". Clearly, the keywords (features) assigned for the two cases will be different too. Therefore, although semi-supervised algorithms have the potential to produce personalized clusters, they have not been explored for this purpose.

In this paper, we conduct a user study to verify whether semi-supervised clustering algorithms can still produce better quality of clusters when human users are asked to perform document supervision and feature supervision than unsupervised clustering without any supervision. At the same time, we explore the semi-supervised algorithms to produce personalized clusters for individual users when combined with their own user input. We develop an interactive interface to help users to group documents and assign keywords for clusters and documents. The interface helps users to create a new cluster, assign a document to an existing cluster, move a document from one cluster to another, merge two clusters, remove assigned documents and existing clusters. Thirty-two participants (users, used exchangeable) are recruited to label 80 out of the 580 documents (academic papers). The 80 papers are selected by an active recommender described in Section 2.3. The papers are generally assigned to three coarse categories assigned by their authors, i.e., software, information systems, and computing methodologies. However, the coarse labels are not used at all in this work, neither for user supervision nor for the evaluation of the algorithms. The participants do not know the actual number of clusters in the document collection and are asked to group the documents based on their own understanding during exploration. In fact, there are no gold-standard labels for this dataset because each user may create any number of sub-clusters within each coarse category. Therefore, we may obtain different sets of clusters of the same 80 documents from each participant, in terms of the number of clusters, the cluster membership of documents and the keywords assigned to clusters. At the same time, they are asked to select the cluster keywords while they are labeling documents. They can also assign keywords to each cluster directly. In order to demonstrate that semi-supervised clustering works with a small amount of user input, only the first few assigned documents (1 to 6) to each cluster are used as document supervision input (see details in Section 3.3). At the same time, only keywords associated with those documents or directly assigned to each cluster are used as feature supervision input. All 580 documents are clustered and the algorithms are evaluated based on the clusters of the 80 documents manually organized by each participant.

In summary, our contributions are: (1) We design and test useful operations and text visualization methods to help users to group documents, which should be included in supervision interface for document management software. We demonstrate that selecting keywords during assigning documents takes little time using the designed interface and operations. (2) We observe that different users group the same document collection differently, i.e., the number of clusters, the cluster memberships of documents, and the assigned keywords. In addition, we observe that a user's organization of a document collection changes over time. Therefore, clustering algorithms which accommodate personalization should be employed. (3) We show that semi-supervised clustering algorithms with a small amount of user input can produce personalized clusters and verify that semi-supervised clustering algorithms can still produce better quality of clusters with (noisy) user input than unsupervised clustering. (4) We demonstrate that assigning keywords for clusters can help clustering algorithms to organize documents better matching user's point of view than any single supervision, i.e., labeling only documents or only features.

The rest of this paper is organized as follows. In Section 2, we present the underlying clustering algorithms, propose an active learning framework to recommend documents for the user to label, and describe the design and components of the interactive user interface to collect user input. Details of the experiments and evaluations are given in Section 3. In this section, we present and discuss the results and observations from our user study. In section 4, we describe the related work. We conclude with a discussion of the implications of this work and the opportunities for further investigations in Section 5.

2. METHODOLOGY

In this section, we first introduce clustering algorithms we use to demonstrate and verify the usefulness of user input, i.e., the unsupervised clustering algorithm KMeans and semi-supervised clustering algorithm $DualSeededK$Means. Then, we briefly describe the active learning method we use to recommend documents for user supervision. Finally, we present the interactive user interface we use to collect user input through document supervision and feature supervision.

2.1 Unsupervised KMeans

KMeans is a clustering algorithm based on iterative assignments of data points to clusters and partitions a dataset into K clusters so that the average squared distance between the data points and the closest cluster centers are locally minimized. For a dataset with data points $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}, x_i \in \mathbb{R}^d$, KMeans algorithm generates K clusters $\{\mathcal{X}_l\}_{l=1}^K$ of $\mathcal{X}$ so that the objective function

$$J = \sum_{l=1}^{K} \sum_{x_i \in \mathcal{X}_l} ||x_i - \mu_l||^2 \qquad (1)$$

is locally minimized based on the initial centers selected, where $\{\mu_1, \mu_2, \ldots, \mu_K\}$ represent the centers of the K clusters.

2.2 Semi-supervised $DualSeededK$Means

$DualSeededK$Means [10, 11] is a semi-supervised algorithm which can incorporate user input from both document supervision and feature supervision. It transforms user input from document supervision into document seeding [2, 10] using clusters derived from labeled documents and user input from feature supervision into feature seeding using a Feature-Vote-Model or Feature-Generative-Model [10]. Finally, it combines document seeding and feature seeding using the linear opinion pool [16]. $DualSeededK$Means is so general framework that it becomes $DocumentSeededK$-Means without feature supervision and $FeatureSeededK$-Means without feature supervision. In fact, $DualSeededK$-Means without any supervision is equivalent to unsupervised KMeans.

2.3 Active Document Recommendation for User Supervision

Since user supervision is labor-intensive, an active learning scheme is designed to recommend the most potentially informative documents for the user to label, i.e., assigning the documents to a cluster. Our algorithm is an adapted version of the explore-consolidate framework [3] to the situation when the number of clusters K is not predefined. In the original explore-consolidate framework described in [3], there are two steps to construct the cluster structure, i.e., “explore” and “consolidate”. In addition, an oracle is used and the number of clusters K is assumed to be known. In each iteration of the “explore” step, a document farthest from the assigned documents is selected using a farthest-first traversal scheme. Then, the document is either assigned to an existing cluster or a new cluster. This step stops after K clusters are created. In each iteration of the “consolidate” step, a document is randomly selected and assigned to one of the existing K clusters. The purpose of this step is to consolidate the cluster structure faster because all clusters exist and there is no need to search for the farthest document. However, it is not directly applicable to our work because human users create clusters according to their own understanding of the document collection and different users may create different numbers of clusters (unknown K). Therefore, we do not know when the “consolidate” step should start. In the adapted version, the “explore” and “consolidate” steps are interleaved. One iteration of the “consolidate” step is performed after every s (4 in this paper) iterations of the “explore” step. However, instead of random selection, a document closest to the smallest cluster is selected in the “consolidate” step. The main goal is to have balanced clusters and avoid having too many small clusters.

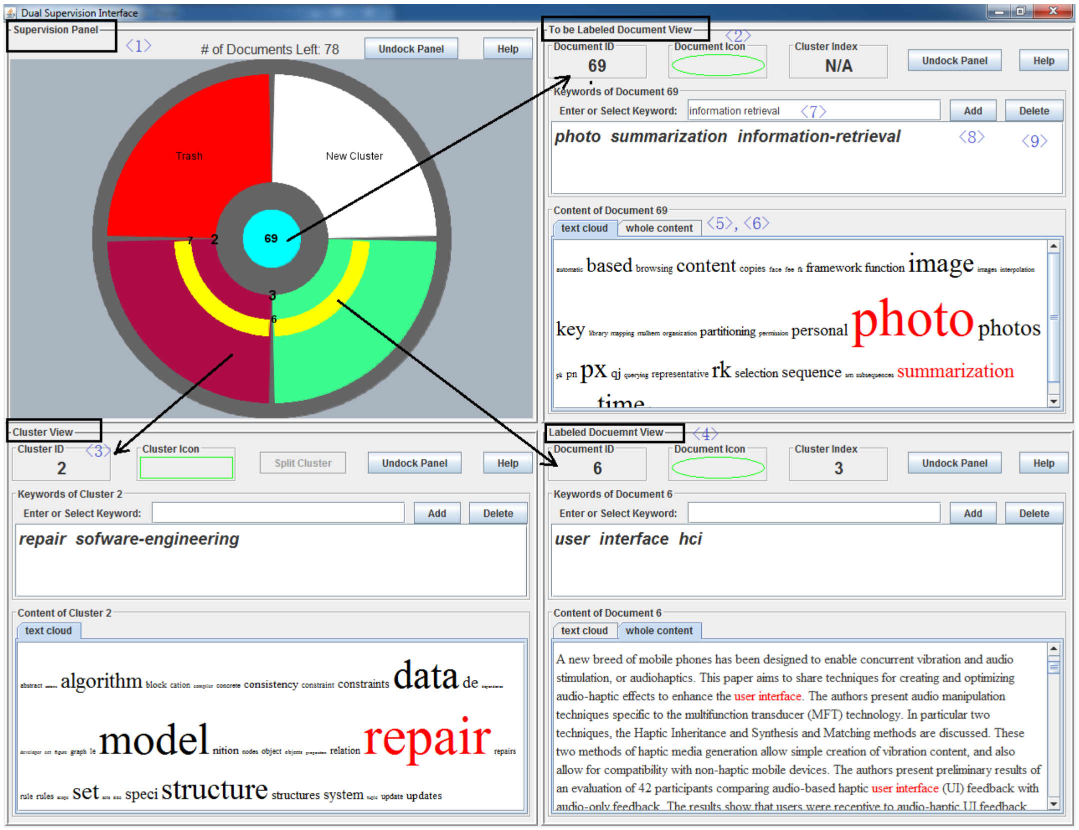
2.4 User Interface for User Supervision

As we mentioned in Section 1, we have two types of user supervision, namely, document supervision and feature supervision. Therefore, we need to provide operations in the user interface to support both types of supervision. We also have to provide visualizations of clusters and documents to aid user supervision. As shown in Fig 1, we have four panels in the user supervision interface:

(1) “Supervision Panel” <1>[2]: This panel supports document supervision. The sectors of the outside circle denote the clusters and the inside circle represents the document that needs to be labeled (assigned to a cluster) by the user. The (yellow) slices inside a sector denote the documents assigned to the corresponding cluster. The number inside a circle, at the top left corner of a slice or sector is the document or cluster ID. There are always two auxiliary sectors, “New Cluster” and “Trash”, which are used to create new clusters and remove clusters or documents respectively. The operations provided by this panel include: (a) Create a new cluster: Drag the inside circle or a slice to the “New Cluster” sector. (b) Move a document: Drag a slice from one sector to another. (c) Merge two clusters: Drag a sector to another. (d) Remove a cluster: Drag a sector to the “Trash” sector. (e) Remove (unlabel or unassign) a document: Drag the inside circle or a slice to the “Trash” sector.

(2) “Document To-Be-Labeled Panel” <2>: This panel displays the information of the document denoted by the inside circle in the “supervision panel” and the document ID matches the one in the inside circle. This panel includes two sub-panels to aid users in identifying the topic of the document, i.e., text cloud <5> [13] and the whole content <6> of the document. The user can select a keyword in either sub-panel, i.e., labeling a feature, by double-clicking on the word. After being chosen as a keyword, the word is highlighted in red. If a word is already being highlighted, double-clicking on it removes the highlighting and it is not a keyword any more (unlabeling a feature). The user can add and delete keywords by using the input field <7> and using the corresponding add/delete buttons <8,9> respectively. All

[2] Corresponding Identification number in Fig. 1



Dual Supervision Interface
Supervision Panel
⟨1⟩
Undock Panel
Help
Trash
New Cluster
69
7
2
3
6
To be Labeled Document View
⟨2⟩
Undock Panel
Help
⟨7⟩
Add
Delete
⟨8⟩
⟨9⟩
text cloud
whole content
⟨5⟩, ⟨6⟩
automatic based browsing content copies face fee ft framework function image images interpolation
key library mapping mulhem organization partitioning permission personal photo photos
pk pn px qj querying representative rk selection sequence sm subsequences summarization
time
Cluster View
⟨3⟩
Split Cluster
Undock Panel
Help
Add
Delete
text cloud
abstract algorithm block cation compiler concrete consistency constraint constraints data de
developer figure graph le model nition nodes object objects relation repair repairs
rule rules set speci structure structures system update updates
Labeled Docuemnt View
⟨4⟩
Undock Panel
Help
Add
Delete
text cloud
whole content

# of Documents Left: 78

Document ID: 69

Document Icon

Cluster Index: N/A

Keywords of Document 69

Enter or Select Keyword: information retrieval

photo summarization information-retrieval

Content of Document 69

Cluster ID: 2

Cluster Icon

Keywords of Cluster 2

Enter or Select Keyword:

repair sofware-engineering

Content of Cluster 2

Document ID: 6

Document Icon

Cluster Index: 3

Keywords of Document 6

Enter or Select Keyword:

user interface hci

Content of Document 6

A new breed of mobile phones has been designed to enable concurrent vibration and audio stimulation, or audiohaptics. This paper aims to share techniques for creating and optimizing audio-haptic effects to enhance the user interface. The authors present audio manipulation techniques specific to the multifunction transducer (MFT) technology. In particular two techniques, the Haptic Inheritance and Synthesis and Matching methods are discussed. These two methods of haptic media generation allow simple creation of vibration content, and also allow for compatibility with non-haptic mobile devices. The authors present preliminary results of an evaluation of 42 participants comparing audio-based haptic user interface (UI) feedback with audio-only feedback. The results show that users were receptive to audio-haptic UI feedback

Figure 1: Interface for User Supervision: Document Supervision and Feature Supervision

keywords of this document will be shown in the keyword area of this panel.

(3) "Cluster View Panel" <3>: After the user single-clicks on a sector in the "supervision panel", this panel displays information about the corresponding cluster. This panel is similar to the "Document To-Be-Labeled Panel" except that there is no visualization of the whole content simple because a cluster does not have it. The user can assign keywords using the methods introduced previously. Note that keywords assigned to a document become keywords of its cluster while the keywords directly assigned to a cluster are not connected to any document assigned to it. Keywords assigned into a cluster should describe the topic of the cluster as they are used by *DualSeededK*Means with Feature-Vote-Model or Feature-Generative-Model in Section 2.2.

(4) "Document Labeled Panel" <4>: This pane's layout is the same as the "Document To-Be-Labeled Panel". When the user single-clicks a slice in the "Supervision Panel", the information about the assigned document is shown here. The user can view the topic of the document and revise the keywords assigned to the document.

3. EXPERIMENTS

3.1 Datasets

The dataset we use for the user study is a collection of the 580 academic papers in full text from different areas of computer science. Those papers were manually collected by the authors from the ACM Digital Library. Based on the 1998 ACM Computing Classification System, those papers were assigned to one or more of the following areas by their authors: Software including Software Engineering and Programming Languages, Information Systems and Computing Methodologies. Generally speaking, the categories assigned by paper authors are very coarse and cannot reflect the accurate topics of the papers. In addition, it is not uncommon that one paper is related to multiple topics and can be assigned to multiple categories. Therefore, this dataset is well suited for us to verify whether different users have their own points of view of the same document collection. At the same time, we can demonstrate the usefulness of user supervision for producing personalized organization.

3.2 Evaluation Measures

We use Rand Distance based on Rand Index [18] to compare different users' clusterings (groupings) of the same document collection and determine whether different users have their own points of view, thereby motivating the inclusion of user personalization as a requirement for clustering algorithms. Based on Rand Index, we develop measures of cohesiveness and separation to evaluate the clusters produced by clustering algorithms in comparison with users' manual organizations. In addition, we use Jaccard distance [19] to measure the dissimilarity between the sets of features labeled by different users.

3.2.1 Rand Distance

We assume a document collection $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ and two clusterings of $\mathcal{D}$, i.e., $\mathcal{X} = \{x_1, x_2, \ldots, x_r\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_s\}$, where x_i or y_j is a subset of $\mathcal{D}$. We also have $x_i \cap x_j = \emptyset$ and $\cup_{i=1,\ldots,r} x_i = \mathcal{D}$ where $i, j \in \{1, \ldots, r\}$ and $i \neq j$, and $y_i \cap y_j = \emptyset$ and $\cup_{i=1,\ldots,s} y_i = \mathcal{D}$ where $i, j \in \{1, \ldots, s\}$ and $i \neq j$. We define the following quantities:

- a, the number of pairs of documents that are in the same cluster in $\mathcal{X}$ and $\mathcal{Y}$.
- b, the number of pairs of documents that are in different clusters in $\mathcal{X}$ and $\mathcal{Y}$.
- c, the number of pairs of documents that are in the same cluster in $\mathcal{X}$ but in different clusters in $\mathcal{Y}$.
- d, the number of pairs of documents that are in different clusters in $\mathcal{X}$ but in the same cluster in $\mathcal{Y}$.

The Rand Index, $\mathcal{RI}$, is:

$$\mathcal{RI} = \frac{a+b}{a+b+c+d} \tag{2}$$

and the Rand Distance, $\mathcal{RD}$, is:

$$\mathcal{RD} = 1 - \mathcal{RI} = \frac{c+d}{a+b+c+d} \tag{3}$$

Rand Index and Rand Distance measure the similarity and the dissimilarity between two clusterings respectively.

3.2.2 Cohesiveness, Separation, and F-Measure

The clusters produced by clustering algorithms are evaluated against users' manual organizations of the document collection. Therefore, we do not use the Rand Index, which only computes the similarity between two clusterings. Instead, we develop measures *coh*, *sep*, and F-Measure to evaluate the clusters produced for this user with/without supervision. Those measures treat a user's manual organization as the gold standard. Assuming the gold standard partition $\mathcal{G} = \{g_1, g_2, \ldots, g_k\}$ and a clustering $\mathcal{C}$ produced by a clustering algorithm, we define the following quantities:

- a', the number of pairs of documents that are in the same cluster in $\mathcal{G}$.
- b', the number of pairs of documents that are in the same cluster in $\mathcal{G}$ and $\mathcal{C}$.
- c', the number of pairs of documents that are in different clusters in $\mathcal{G}$.
- d', the number of pairs of documents that are in different clusters in $\mathcal{G}$ and $\mathcal{C}$.

The cohesiveness of $\mathcal{C}$, *coh*, is:

$$coh = \frac{b'}{a'} \tag{4}$$

The separation of $\mathcal{C}$, *sep*, is:

$$sep = \frac{d'}{c'} \tag{5}$$

and finally F-Measure, F, is:

$$F = 2 \times \frac{coh \times sep}{coh + sep} \tag{6}$$

where *coh* measures the cohesiveness of $\mathcal{C}$ while *sep* measures the separation of $\mathcal{C}$, based on a user's manual organization $\mathcal{G}$.

3.2.3 Jaccard Distance

Given two sets $\mathcal{A}$ and $\mathcal{B}$, the Jaccard Index, $\mathcal{JI}$, is:

$$\mathcal{JI} = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \tag{7}$$

and the Jaccard Distance, $\mathcal{JD}$, is:

$$\mathcal{JD} = 1 - \mathcal{JI} = \frac{|\mathcal{A} \cup \mathcal{B}| - |\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \tag{8}$$

Jaccard Index, $\mathcal{JI}$, measures similarity between two sets while Jaccard Distance, $\mathcal{JD}$, measures dissimilarity between two sets. Given two clusterings $\mathcal{X}$ and $\mathcal{Y}$ of a document collection $\mathcal{D}$ (Section 3.2.1), $\mathcal{X}_w = \{x_{w1}, x_{w2}, \ldots, x_{wr}\}$ and $\mathcal{Y}_w = \{y_{w1}, y_{w2}, \ldots, y_{ws}\}$ are the sets of keywords assigned to each cluster by users, i.e., x_{wi} and y_{wj} are the keywords assigned to cluster x_i and y_j respectively. We define two dissimilarity measures between $\mathcal{X}_w$ and $\mathcal{Y}_w$. One measure $\mathcal{JD}_a$ measures dissimilarity between $\mathcal{X}_w$ and $\mathcal{Y}_w$ without consideration of the cluster labels of the assigned keywords, i.e., $\mathcal{A} = \cup_{i=1,2,\ldots,r} x_{wi}$ and $\mathcal{B} = \cup_{j=1,2,\ldots,s} y_{wj}$ in Eq. 8. The other measure $\mathcal{JD}_b$ measures dissimilarity between $\mathcal{X}_w$ and $\mathcal{Y}_w$ with cluster labels considered. $\mathcal{JD}_b$ is defined as:

$$\mathcal{JD}_b = \frac{\sum_{i=1}^{r} \min_{j=1,\ldots,s} \mathcal{JD}(x_{wi}, y_{wj}) + \sum_{j=1}^{s} \min_{i=1,\ldots,r} \mathcal{JD}(y_{wj}, x_{wi})}{r+s} \tag{9}$$

In this measure, we compute the average distance between a cluster and its closest match in the other clustering. A closest match is from the other clustering and has minimum distance from a cluster. In this way, cluster labels are considered when the measure is computed. Note that the closest match relationship is not symmetrical, i.e., with x_{wi}'s closest match being y_{wj}, the closest match of y_{wj} could be x_{wk}, where k might not be same as i.

3.3 Experimental Setup

We recruited thirty-two participants to group 80 of the 580 academic papers in our ACM dataset [3]. These 80 papers are selected by the active learning method presented in Section 2.3 and every participant groups the same 80 papers. The thirty-two participants include 5 female and 27 male graduate students from computer science. At the beginning, the task of the user study is introduced to all participant that there are not given predefined categories and they are asked to group papers based on their own understanding during the exploration of the collection. They are also aware that they need to assign keywords to a document and those keywords will become the cluster keywords automatically after the document is assigned to a cluster. They can also assign and remove keywords to and from clusters directly. Then, they are demonstrated how to group the documents and assign keywords using the software, whose interface is shown in Fig. 1 and then they are given 5 minutes to get familiar with the software. Finally, they are asked to use the software to group the 80 documents. At one time, there is only one document to be labeled (represented by the inside circle). The order of appearance of the

[3] The dataset and manual organizations from all participants will be available from the first author's homepage: `http://www.cs.dal.ca/~yeming/`. This will provide a dataset with multiple (noisy) organizations to evaluate future clustering algorithms.

80 documents depends on the active learning recommendation method. Although all participants group the same 80 documents, the order of documents appear for each user is different.

For all users, we experiment with document supervision consisting of fewer than the full 80 documents a user labels. We only use the first m documents assigned to each cluster, where m ranges from 1 to 6. Documents within each cluster are ordered based on the time they were assigned to the cluster, either when being labeled for the first time or when moved from another cluster. When a cluster B is merged into cluster A, documents in A precede documents in B. At the same time, only keywords associated with those documents selected for document supervision, or directly assigned to each cluster, are used as feature supervision input. Since the order the documents appear for labeling is distinct, the user input from each user for *DualSeededK*Means includes different sets of documents and labeled features. All 580 papers are clustered and the clusterings produced from different algorithms are evaluated based on the 80 user labeled papers using *coh*, *sep*, and F-Measure.

3.4 Results

In this section, we present the user feedback and analyze results from our user study. We performed three kinds of analysis on the following aspects: user behaviors using the interface, personalization of the same document collection, and personalized document clustering with dual supervision.

3.4.1 User Satisfaction with the User Interface

Generally speaking, all participants think they know the topics of document collection well and it is easy to identify the topic of a document and identify the keywords that need to be assigned into a cluster after they manually organize the 80 documents recommended by the active recommender. They also indicate that the operations to assign and move a document, delete and merge clusters are easy to use. However, only one participant used the operation "split a cluster" since others did not realize its existence. They would like to use the software to organize their personal library of papers if the system is with proper documentation and agree that the system can help them to organize their papers better. A few interesting points we find out are:

- Twenty-nine participants think assigning keywords only takes a little time (less than 10 seconds) while only three of them indicate that it takes some time (more than 10 second but less than 1 minute). No one thinks it takes much time (more than 1 minute).
- All participants except one think that the whole content is more useful than text cloud in identifying the topic of a document. This point can also be verified by Table 5, which shows that about 70% keywords are labeled in the whole content. It is surprising since we expected that text cloud would be more helpful. One of the possible explanations is that we used single words for text cloud and multiple-word phrases could have made text cloud more useful.
- All participants review the topic of an existing cluster through keywords instead of reading a document assigned to this cluster. Therefore, it is very important to assign meaningful and correct keywords to a cluster from the beginning.

Table 1: Definitions of Operations

Name	Definition
Add	Assign a document into a cluster
Move	Move a document to another cluster
Delete	Remove a cluster
Merge	Merge two clusters
Label	Add a keyword by double-clicking
Unlabel	Remove a keyword by double-clicking
AddButton	Add a keyword through Add Button
DelButton	Remove a keyword through Delete Button

Table 2: Statistics of # of Clusters Created, Assigned Documents and Keywords

Name	Range	AVG	MED
# of Clusters	4– 9	6.34	6
Assigned Documents	68–80	76.47	77
Assigned Keywords / Cluster	1–26	9.09	8

In addition, many participants suggest that they would like to have more functionality such as searching documents by words. They also suggest that we might add a spell checker for the keywords they enter. More specifically, some participants like to have all assigned documents with a keyword within a cluster highlighted when the keyword of that cluster is selected.

We present a few excerpts from users' feedback to support our claims:

- "The pie visualization [4] is very easy to use after practicing on it."
- "I really liked the drag and drop feature, which has made the system very easy to use."
- " 'Split A Cluster' helped me when I by mistake merged two clusters together."
- "I like the typing keywords feature because it allows me to generalize or be more specific about keywords without being constrained to a predefined list."
- "I found text clouds less useful than I expected."
- "It should be useful to have cluster or document keywords when the mouse hovers it in the supervision panel."
- "When a document or cluster is selected, I would expect this cluster or document was somehow highlighted in the supervision panel. Without it, it is not easy to move a document from one cluster to another."
- "I'd like to have the search (CTRL+F) function and a spell checker."
- ...

3.4.2 User Behaviors

We analyze operations defined in Table 1 which users use during grouping documents and assigning keywords so we can identify the most useful ones that should be included in

[4] The "Supervision Panel"

Table 3: Statistics of Operations Users Use

Operation	MIN	AVG	MED	MAX
Add	80	80.5	80	92
Move	0	4.31	3	22
Delete	0	3.03	2	10
Merge	0	1.84	1	10
Label	5	72.00	54	205
Unlabel	0	4.15	2	18
AddButton	0	12.53	9	80
DelButton	0	9.28	7	36

Table 4: Keywords Assigned through Documents or Directly

	Through Doc		Directly		Total
Label	68.03	91.13%	3.97	8.87%	72.00
AddButton	9.00		3.53		12.53
Unlabel	3.78	51.90%	0.37	48.10%	4.15
DelButton	3.19		6.09		9.28

Table 5: Keywords Assigned through Text Cloud or Whole Content

Name	Text Cloud		Whole Content		Total
Label	23.03	31.97%	48.97	68.01%	72.00
Unlabel	1.37	33.01%	2.78	66.99%	4.15

Table 6: Frequency of # of Clusters Created

#	4	5	6	7	8	9	Total
Frequency	2	7	11	6	2	4	32
Percentage	6.25	21.88	34.38	18.75	6.25	12.5	100

future interface design. We also present the analysis of the text visualization methods used in the user interface.

A document is considered as "assigned" after it is placed into an existing or a newly created cluster. Otherwise, it is considered as "unassigned", i.e., the document is placed into the "trash" cluster. The average of assigned documents out of the 80 documents is 76.47 (Table 2). Therefore, users know topics of most documents recommended by the active recommender. The most unassigned documents by a user is 12, which is 15% of all documents. However, most users only have less than 4 documents not assigned to any cluster. In fact, some users put a few documents into trash clusters at the beginning. Later on, those documents are retrieved and assigned into an existing cluster. Oracles in previous work could not simulate this behavior, in which not all recommended documents are assigned at the beginning and users can have chance to put some documents on hold and cluster them later. In addition, users are able to assign at least a few keywords for each cluster and generally assign about 9 distinct keywords for each cluster (Table 2).

In Table 3, we display the minimum, average, median, and maximum times users use each operation. The fact that we have 92 (more than 80) add operations indicates that some documents are moved from an existing cluster to create a new cluster. It is also not uncommon that a user move a document from one cluster to another, delete an existing cluster and merge two existing clusters. In addition, users also assign plenty of keywords for clusters. At the same time, many keywords are removed after they are assigned. On one hand, keywords are assigned through double-clicking more often than using add keywords buttons. That's mostly due to the fact users can assign the keywords during reading a document. However, some keywords have to be assigned or removed through add or delete keyword buttons because these keywords do not exist or are difficult to find in any document. On the other hand, users remove keywords mainly through the delete button. That's because users normally clean the keywords for a cluster using delete button at the end of the manual organization so that the keywords left can represent the topic of the cluster well. All frequent use of those operations verify that a user can change his perception of the document collection while exploring the document collection. Therefore, clustering software should enable users to change the existing cluster structures.

Next, we analyze user behaviors on assigning keywords. A user can assign keywords to a document and the keywords associated with the document are assigned to a cluster automatically after the document is assigned into the cluster. A user can also assign keywords to a cluster directly. In addition, we are interested in which visualization method users use most often. From Table 4, we observe that most keywords are assigned through documents while a small percentage is assigned into clusters directly. Although most participants assigned keywords primarily by double-clicking and rarely by using "AddButton", one participant only used "AddButton" because he said the keywords that he came up with could best reflect his perception of the collection. With regards to keyword removals, users removed keywords equally often through documents and directly from clusters. It is observed that keyword removals through documents mostly take place when users read documents and try to learn its topic while users remove keywords from clusters directly after they finish the manual organizations and want to clean the keywords which represent the topics of clusters. From Table 5, we observe that more than two-thirds of the keywords assigned through double clicking are selected using whole content of document. This fact is consistent with the user feedback that the whole content is more helpful in discerning the topic of the document than the text cloud. However, text cloud is still useful for assigning keywords since it has fewer words than whole content and easier to find the word to be assigned. As some users indicated in the post-study questionnaires, text cloud is useful to have a general idea about the document but the whole content helps to find the exact topic of the document.

3.4.3 Personalization

We compare different groupings of the same 80 papers from all participants in terms of both documents and keywords assigned. We also compare the groupings of the same 80 papers from the same users at different times to see whether the same users have different views of the same collection over time.

First, users create different numbers of clusters based on their own understanding although all numbers are between

Table 7: Statistics of User Manual Organizations

	MIN	AVG	MED	MAX
$\mathcal{RD}$	0.1308	0.2483	0.2455	0.3817
$\mathcal{JD}_a$	0.6346	0.8632	0.8677	1.0
$\mathcal{JD}_b$	0.6483	0.9007	0.9082	1.0

Table 8: Manual Organizations by Five Users at Different Times

	P_1	P_2	P_3	P_4	P_5	AVG
$\mathcal{RD}$	0.2202	0.073	0.1652	0.1323	0.1647	0.1511
$\mathcal{JD}_a$	0.5574	0.4444	0.82	0.8684	0.6667	0.6714
$\mathcal{JD}_b$	0.6072	0.4348	0.8396	0.8269	0.7498	0.6917

4 and 9 (Table 2). More specifically, about 80% of the participants created 5, 6, or 7 clusters while others created 4, 8, 9 clusters (Table 6). The frequency of the cluster numbers are close to uniform distribution among 5, 6, 7 (the more frequent cluster numbers, about 74%) and among 4, 8, 9 (the less frequent ones, about 36%) respectively. Therefore, users tend to create different numbers of clusters. Later on, we will observe that different participants have distinct clusters regardless of whether the cluster numbers are the same or not.

We present the minimum, average, median and maximum Rand Distance and Jaccard Distance between organizations of all user pairs of clusterings in Table 7. The average Rand Distance between the user pair organizations is about 0.25. If different users create similar partitions of the same document collection, we would expect that the average Rand Distance is close to 0. Therefore, the Rand Distance 0.25 shows that there is substantial disagreement between different users and distinct clusters were created. In addition, the Jaccard Distances in terms of labeled keywords indicate even more disagreement between different users (average distance about 0.90). That is because there is normally a much bigger word vocabulary than the number of documents and many different keywords can be used to identify the same cluster topic, i.e., completely different keyword sets can be used for the same topic. $\mathcal{JD}_b$ shows a higher disagreement than $\mathcal{JD}_a$ because $\mathcal{JD}_b$ considers the cluster label of keywords while $\mathcal{JD}_a$ does not. Therefore, it confirms our conjecture that different users have different points of view of the same document collection.

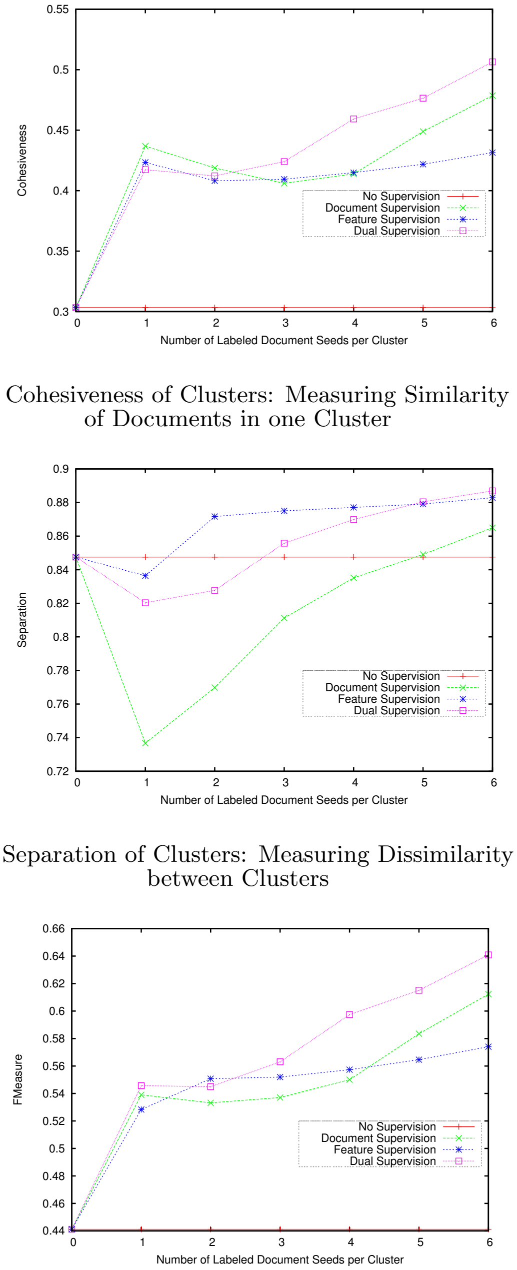
Finally, we compare manual organizations by the same users but at different times. In our user study, we asked the same five users to organize the same document collection again one week after their first participation. Generally speaking, the two organizations are still distinct from one another although they are closer when compared to organizations from different users (Table 7 and Table 8). For example, the average Rand Distance between user pair's manual organizations is 0.25, while from the same user is 0.15.

3.4.4 Document Clustering with User Supervision

Semi-supervised clustering algorithms have been proved able to improve clustering performance over unsupervised peer algorithms using oracles [7, 8, 9, 10]. Since oracles used in previous work are assumed to give "correct" answer all the time, our purpose here is to verify that document clustering with human's noisy supervision can still produce more consistent clusters with user's manual organi-

Table 9: Rand Distances Between Clusterings Produced by Each Algorithm for Different Users

Name	No Supervision	Document	Feature	Dual
$\mathcal{RD}$	0.080	0.2540	0.1711	0.2226



(a) Cohesiveness of Clusters: Measuring Similarity of Documents in one Cluster

(b) Separation of Clusters: Measuring Dissimilarity between Clusters

(c) F-Measure of Clusters: Measuring both Cohesiveness and Separation

Figure 2: Performance of clustering algorithms with no supervision, document supervision, feature supervision and dual supervision

zation than unsupervised clustering techniques. Instead of using a single universal ground truth as in previous work, each user has his own ground truth in our case. Therefore, we also explore whether document clustering with user supervision can produce personalize clusters. We present the results of clustering algorithms initialized by a few (1 to 6) documents and the keywords labeled in those documents or assigned directly into the clusters. We use *coh*, *sep*, and F-Measure to quantify the consistency of the computed clustering with the user's manual organization. From Fig. 2(a) and Fig. 2(b), we can tell that document supervision is able to group similar documents together while feature supervision is better at separating dissimilar documents. A small number of labeled documents appears to lead to unbalanced clusters with high cohesiveness (e.g. *coh* is 1 when all documents are place into the a single cluster), since the labeled documents can not represent the cluster structure well. We observe this behavior in Fig. 2(b). A small number of keywords assigned to clusters (through the labeled documents or directly by the user) appears to lead to more balanced clusters with a better representation of cluster structures, as seen in Fig. 2(b). Especially, document supervision can not separate dissimilar documents very well into different clusters when less than 4 labeled documents per cluster are provided (Fig. 2(b)). Since the keywords (features) assigned by users are representative of the clusters, feature supervision can provide good performance in terms of both cohesiveness and separation. Dual supervision, the combination of document supervision and feature supervision, can generally produce clusters better matching user's expectation (Fig. 2) [5]. Since assigning keywords is efficient for users, it is worth the effort to improve the clustering performance. In addition, the performance of the clustering algorithms improves with more labeled documents (and more assigned keywords) for initialization. It is easily understandable since more documents and/or keywords can represent the cluster structures better.

We also investigate how consistent the clusterings are between different participants. In Table 7, we compute and display the average Rand Distances between clusterings generated by the same algorithm for different users with same type and amount of supervision. Like the manual organizations from each user (Table 7), the clusterings produced with user's input are also distinct from each other (Table 9). In Table 9, the average Rand Distance between clusterings produced with document supervision and dual supervision with 4 documents and associated keywords is about 0.23, which is very close to the average Distance between different users' manual organization (about 0.25 in Table 7). On the other hand, the Rand Distance between clusterings produced without supervision is only 0.08. Therefore, clusterings obtained via semi-supervised clustering with user supervision are highly depended on user input, and therefore they can be viewed as personalized clusterings.

4. RELATED WORK

Traditional semi-supervised clustering techniques normally employ user supervision in the form of document-level constraints. The document constraints are generally used to modify the loss functions [4], initialize the cluster centers [2], learn adaptive distance metrics [22], and project high dimensional feature space to lower dimensional subspaces [20]. Recently, an alternative form of user supervision such as labeling features has been explored to aid semi-supervised clustering algorithms [8, 9, 10, 12]. Some work [8] uses only labeled words to guide clustering algorithms while others [9, 10, 12] integrate both labeled documents and words into a unified framework. However, these works are either evaluated based on oracles or no formal user study is performed. Drucker et al. [6] propose to use adaptive machine learning recommendations to help users group large numbers of documents faster. They did a formal user study by recruiting thirty-two participants to group the same document collection. Then, they asked a pool of 161 raters to rate the clusterings produced from the 32 participants. The user study demonstrates that clusters produced with the help of the adaptive machine learning method are significantly better than clusters automatically created. However, their work did not evaluate user's personalization of the same document collection since the generated clusters are not evaluated by users who provided the supervision. Feature supervision was also used to improve the performance of classification algorithms, such as using the labeled features for each class to constrain the probabilistic model estimation [5], making use of feature feedback with support vector machine [17], and creating pseudo-instances using the labeled features for each class [15], etc. However, classification methods assume there are pre-defined categories to which users can assign documents or keywords. In document clustering, users have to form their perception of the document collection during exploration. In addition, oracles were employed to evaluate the proposed algorithms and no formal user study is conducted in those works. Personalization has been explored in the clustering of search results [1] and search engine queries [14].

5. CONCLUSIONS AND FUTURE WORK

We recruited thirty-two participants to organize the same document collection. We analyzed users' behaviors during their manual organization. The analysis shows that users can easily find the keywords to assign to a cluster based on the whole content of the documents and it is efficient according to users' feedback. Instead of only assigning keywords existing in the documents, users also like to come up with phrases to describe the topics of clusters. By comparing all groupings from all participants, we find that each user has his own perception of the document collection and a clustering algorithm with user supervision is required to produce personalized clusters, which better reflect his point of view. At the same time, we confirm that previously proposed semi-supervised document clustering algorithms can produce personalized clusters with a small amount of user input even if it is noisy. It is also demonstrated that the same user can change his perception of the documents over time. Therefore, operations such as moving a document between clusters and merging two clusters should be available in software for document clustering. We also find that text cloud with single words is less useful than the full text for users to grasp the topic of a document.

Since text cloud with single words is not as helpful, it is worth to further investigate text cloud with multiple-word terms. According to users' feedback, functions such as searching documents by words, retrieving documents in a cluster with a specific keyword of that cluster, and a spell

[5] Two-tailed paired t-test with p = 0.05.

checker should be added to the user interface in the future. Since a user changes his perception of the document collection during exploration, the software should be able to interleave user supervision and clustering, i.e., the user should be able to make adjustments of documents and features after intermediate clusters are obtained, and then the clustering procedure is repeated with the updated user input. Other future work directions include enabling users to create hierarchical clusters with the user interface and allowing soft clustering, namely, a document to be assigned to multiple clusters.

References

[1] D.C. Anastasiu, B.J. Gao, and D. Buttler. Clusteringwiki: personalized and collaborative clustering of search results. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1263–1264, New York, NY, USA, 2011. ACM.

[2] S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *International Conference on Machine Learning*, pages 19–26, 2002.

[3] S. Basu, A. Banerjee, and R.J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 333–344, 2004.

[4] S. Basu, M. Bilenko, and R.J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 59–68. ACM, 2004.

[5] G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602. ACM, 2008.

[6] S.M. Drucker, D. Fisher, and S. Basu. Helping Users Sort Faster with Adaptive Machine Learning Recommendations. In *Proceedings of the 13th International conference on Human-Computer Interaction*, pages 187–203, 2011.

[7] Y. Hu, E.E. Milios, and J. Blustein. Interactive Document Clustering Using Iterative Class-Based Feature Selection. Technical report, CS-2010-04, Faculty of Computer Science, Dalhousie University, Canada, 2010.

[8] Y. Hu, E.E. Milios, and J. Blustein. Interactive feature selection for document clustering. In *Proceedings of the 26th Symposium On Applied Computing, On Track "Information Access and Retrieval"*, pages 1148–1155. ACM Special Interest Group on Applied Computing, 2011.

[9] Y. Hu, E.E. Milios, and J. Blustein. Enhancing Semi-supervised Document Clustering with Feature Supervision. In *Proceedings of the 27th ACM Symposium Applied Computing, On Track "Information Access and Retrieval"*, pages 950–957. ACM, 2012.

[10] Y. Hu, E.E. Milios, and J. Blustein. Semi-supervised Document Clustering with Dual Supervision through Seeding. In *Proceedings of the 27th ACM Symposium Applied Computing, On Track "Data Mining"*, pages 463–470. ACM, 2012.

[11] Y. Hu, E.E. Milios, and J. Blustein. A unified framework for document clustering with dual supervision. *ACM Applied Computing Review*, 12(2), 2012.

[12] Y. Huang and T.M. Mitchell. Text clustering with extended user feedback. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 420. ACM, 2006.

[13] J. Lamantia. Text Clouds: A New Form of Tag Cloud? http://www.joelamantia.com/tag-clouds/text-clouds-a-new-form-of-tag-cloud, 2007. Accessed on April 12, 2012.

[14] K.W. Leung, W. Ng, and D. Lee. Personalized concept-based clustering of search engine queries. *IEEE Trans. on Knowl. and Data Eng.*, 20(11): 1505–1518, 2008.

[15] B. Liu, X. Li, W.S. Lee, and P.S. Yu. Text classification by labeling words. In *Proceedings of the National Conference on Artificial Intelligence*, pages 425–430, 2004.

[16] P. Melville, W. Gryc, and R.D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284, 2009.

[17] H. Raghavan, O. Madani, and R. Jones. Interactive feature selection. In *Proceedings of IJCAI 05: The 19th International Joint Conference on Artificial Intelligence*, pages 841–846, 2005.

[18] W.M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, pages 846–850, 1971.

[19] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, 2005.

[20] W. Tang, H. Xiong, S. Zhong, and J. Wu. Enhancing semi-supervised clustering: a feature projection perspective. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 707–716. ACM, 2007.

[21] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584, 2001.

[22] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems*, pages 521–528, 2003.