A Feature-Enhanced Ranking-Based Classifier for Multimodal Data and Heterogeneous Information Networks

Scott Deeann Chen[1], Ying-Yu Chen[1], Jiawei Han[2], and Pierre Moulin[1]

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
[2]Department of Computer Science, University of Illinois at Urbana-Champaign

***Abstract*—We propose a heterogeneous information network mining algorithm: feature-enhanced RankClass (F-RankClass). F-RankClass extends RankClass to a unified classification framework that can be applied to binary or multiclass classification of unimodal or multimodal data. We experimented on a multimodal document dataset, 2008/9 Wikipedia Selection for Schools. For unimodal classification, F-RankClass is compared to support vector machines (SVMs). F-RankClass provides improvements up to 27.3% on the Wikipedia dataset. For multimodal document classification, F-RankClass shows improvements up to 19.7% in accuracy when compared to SVM-based meta-classifiers. We also study 1) how the structure of the network and 2) how the choice of parameters affect the classification results.**

***Index Terms*—classification, multimodal, heterogeneous information network, ranking**

I. Introduction

Multimodal data mining has been actively researched over the past decades. Mining algorithms discover and use relationships between modalities for content analysis. A wide variety of applications are found in the literature, such as video-audio, video-captions, text-image, image-captions analyses. However, most multimodal mining and fusion algorithms focus on audio and video analysis [1] but not on multimodal document mining, which analyzes text and images in documents.

Multimodal documents, which typically contain both text and images, are becoming more and more ubiquitous with the advance of information technology. Text-based mining techniques have been extremely successful in a wide variety of applications [2]. However, images, the other mode that often accompanies text, has not been considered much. To the best of our knowledge, only little work is found on multimodal document indexing, retrieval [3], and classification [4] [5]. Some image mining frameworks also exploit text-image relationships and use text to aid image mining, but most of them only involve image captions [6] [7].

Our work focuses on overcoming the hurdles in multimodal document mining. The framework we develop is however not limited to multimodal documents but is applicable to other multimodal content analyses as well. The challenges are discussed as follows. First of all, as pointed out by Chen *et al.* [5], the number of images per document differs across documents. There are no straightforward methods to obtain a fixed-dimension feature vector from a document. Therefore, previous developments on multimodal fusion frameworks, which requires a fixed feature dimension, may not be directly applicable on multimodal documents.

Second, images of multimodal documents are by nature more diverse and may not directly provide the most relevant information. This is not only true for user-generated documents, such as blogs, personal websites, etc., but also for more formal documents. Materials put together are not always consistently relevant. In particular, the relationship between the topic and an image in the document is not always clear or may not even exist, as in the case of advertisements on webpages.

Third, multimodal fusion frameworks [1] fuse information from different modalities through 1) *early-fusion*, concatenating features from different modalities, or 2) *late-fusion*, using first-stage analysis units to bring different modalities to a common ground. Early-fusion does not preserve the multimodal structure, and, therefore, structure information is not considered in further analyses. On the other hand, late-fusion has less information for further analyses due to the data processing inequality. The ideal case is to preserve both information of features and multimodal structure.

To that end, we propose a heterogeneous information network mining framework: **feature-enhanced** RankClass (**F-RankClass**). The original RankClass performs ranking and classification at the same time on heterogeneous information networks. F-RankClass extends RankClass [8] to a more general framework that is applicable to most classification problems. In this paper, we construct heterogeneous information networks from multimodal documents and perform classification according to network structures.

The merits of the F-RankClass framework are: 1) F-RankClass does not require features extracted from multimodal data to have a fixed number of dimension. Only extra links between nodes representing features and data objects are added. 2) Inheriting benefits from RankClass, F-RankClass identifies irrelevant objects by ranking, and thus improves classification accuracy. 3) All information, including features and multimodal structures, are encoded in the heterogeneous information network without further information loss. 4) F-RankClass provides a unified framework for both binary and multiclass classification of multimodal and unimodal classification without further modification.

The rest of this paper is organized as follows. Section 2 reviews related work in heterogeneous information network mining and multimodal document classification. Section 3 describes the F-RankClass framework and the application on multimodal document classification. Section 4 presents experimental results on unimodal and multimodal classification

The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA) and U.S. National Science Foundation grants CNS-0931975, IIS-1017362, and IIS-1320617.

problems. Section 5 summarizes and concludes the paper.

II. Related Work and Background

A. Heterogeneous Network Mining

Heterogeneous information network mining algorithms gained a lot of attention recently. PopRank [9] considered different types of links between nodes to rank web-objects. NetClus [10] used a star schema to represent heterogeneous data and considered ranking information within each object type. More recently, Ji *et al.* [8] proposed RankClass, which performs ranking and classification at the same time. Ranking and classification enhance each other in RankClass. Ranking discovers important representative examples for the classifier to learn from, and classification identifies objects of the same class so that within-class ranking becomes more meaningful. More details on RankClass are discussed in Section III-B.

B. Multimodal Document Classification

Joint text and image document classification was first studied by Shatkay *et al.* [4] with a late-fusion framework that fuses base-classifier outputs through a simple OR function. Chen *et al.* used a late-fusion scheme, Support Vector Machine (SVM)-based meta-classifiers [5], and experimented on 2008/9 Wikipedia Selection for Schools [11]. However, no cross-modality relationship was considered during the feature extraction and feature fusion procedure in either [4] or [5].

III. Proposed Mining Framework

A. Heterogeneous Information Networks

We follow the terminology from RankClass [8]. A *heterogeneous information network* consists of multiple types of data objects and links connecting different types of data objects. A heterogeneous information network is represented by a graph $G = (V, E, W)$, where V is a set of data object nodes, E is a set of edges linking data object nodes, and W is a set of edge weights. The node set V is the union of nodes of all object types, i.e., $V = \bigcup_{i=1}^{m} X_i$, where m is the number of object types and X_i is the set of nodes of type i. In a heterogeneous information network, we have $m \geq 2$ types of data objects. We denote by $x_{ip} \in X_i$ the p^{th} node of type i, by $(x_{ip}, x_{jq}) \in E$ an edge linking x_{ip} and x_{jq}, and by $w_{x_{ip},x_{jq}} \in \mathbb{R}^+ \cup \{0\}$ the weight associated with (x_{ip}, x_{jq}).

We represent a subgraph, containing X_i and X_j and the edges between, by a relation matrix $R_{ij} \in \mathbb{R}^{n_i \times n_j}$, where n_i and n_j are the numbers of objects of types i and j, respectively. The p^{th} row and q^{th} column element in R_{ij}, $R_{ij,pq}$, encodes the edge weight $w_{x_{ip},x_{jq}}$. The weight is set to zero if the corresponding edge is absent. Then, we associate each relation matrix R_{ij} with a normalized relation matrix S_{ij}. The normalization is defined as follows:

$$S_{ij} = D_{ij}^{-\frac{1}{2}} R_{ij} D_{ji}^{-\frac{1}{2}},\ i, j \in \{1, \ldots, m\},$$

where $D_{ij} \in \mathbb{R}^{n_i \times n_i}$ and $D_{ji} \in \mathbb{R}^{n_j \times n_j}$ are diagonal matrices. The p^{th} element on the diagonal of D_{ij} is set to be the sum of the p^{th} row of R_{ij}. The q^{th} element on the diagonal of D_{ji} is set to be the sum of the q^{th} column of R_{ij}.

B. Review of RankClass

RankClass classify examples by propagating class information via edges in a heterogeneous information network. A ranking distribution $P(x_{ip}|X_i, k) : X_i \mapsto \mathbb{R}$ is defined over all nodes of type i for the object type X_i and a class k. The larger $P(x_{ip}|X_i, k)$ is, the higher the rank of x_{ip} is among all objects of type i. The ranking distribution is updated iteratively in two steps: 1) propagating class information between adjacent nodes, and 2) adjusting the network to favor within-class ranking.

We first initialize distributions from training data as follows:

$$P(x_{ip}|X_i, k)^0 = \begin{cases} \frac{1}{l_{ik}} & \text{if } x_{ip} \text{ has label of } k, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where l_{ik} is the number of type i training data of class k. We then update $P(x_{ip}|X_i, k)$ iteratively. Step one propagates class information from a node to its neighbors through the following update equation:

$$P(x_{ip}|X_i, k)^t \propto \sum_{j=1}^{m} \lambda_{ij} S_{ij,pq} P(x_{jq}|X_j, k)^{t-1} + \alpha_i P(x_{ip}|X_i, k)^0, \quad (2)$$

where $P(\cdot)^t$ denotes a distribution in the t^{th} iteration, $P(\cdot)^0$ denotes the initial ranking distribution induced by training data, $\lambda_{ij} \in [0, 1]$ controls the amount of information flowing between objects of type i and j, and $\alpha_i \in [0, 1]$ controls the amount of information flowing from the training data. The first term is a weighted sum of ranking probability of the neighbors of x_{ip}. The second term introduces the distribution induced by the training data.

Step two adjusts the network to favor within-class ranking by increasing edge weights between objects that are highly ranked in the same class, through the following update equation:

$$R_{ij,pq}^t(k) = R_{ij,pq} \times \left(r(t) + \sqrt{\frac{P(x_{ip}|X_i, k)^t}{\max_p P(x_{ip}|X_i, k)^t} \frac{P(x_{jq}|X_j, k)^t}{\max_q P(x_{jq}|X_j, k)^t}} \right), \quad (3)$$

where $r(t)$ is a positive number to avoid edge weights dropping to zero in the first several iterations and is set to $r(t) = \frac{1}{2^t}$. This operation allows weights of edges connecting highly ranked objects of the same class to increase and weights of edges connecting less representative objects to decrease.

Finally, each node is assigned the class of the highest posterior class probability, $P(k|x_{ip}, X_i) \propto P(x_{ip}|X_i, k)P(k|X_i)$. The expectation-maximization (EM) algorithm is used to estimate $P(k|X_i)$ through the following equations:

$$P(k|x_{ip}, X_i)^t \propto P(x_{ip}|X_i, k)P(k|X_i)^t \quad (4)$$

$$P(k|X_i)^t = \frac{\sum_{p=1}^{n_i} P(k|x_{ip}, X_i)^t}{n_i}. \quad (5)$$

C. F-RankClass

F-RankClass is a classification framework based on RankClass. F-RankClass relates each data object with features extracted from each data object in addition to binary-weighted

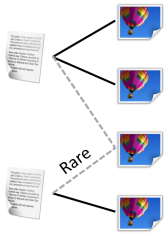


Fig. 1: A path between two documents exists only when the same image appear in both documents. This situation is possible but rare.

edges as in RankClass. We view features as a type of node and each feature as a node. A data object node is connected to a feature node when the feature value of the data object is non-zero. The edge weight is a real positive number set to be the feature value. An offset is added to features with negative values to prevent negative edge weights.

In F-RankClass, objects do not have to be directly related but have to share common features. Hence, F-RankClass can be applied to a whole new spectrum of applications compared to RankClass. For example, if we want to build a graph from a multimodal document dataset. RankClass only links text and images that coexist in a document, as in Fig. 1. We can mine a graph by RankClass only when the graph is not disjoint; however, two sets of documents without common images are disjoint. On the other hand, F-RankClass utilizes term-frequency features and represents a term by a node. A term node is connected to a text node with edge weights equal to the term-frequency. Since most pairs of documents has common words, the graph will not be disjoint with high probability.

Formally speaking, given a graph with m types of data objects $X_1, \ldots, X_m$, we define another k types of nodes $X_{m+1}, \ldots, X_{m+k}$ representing different types of features that can be extracted from the data objects. We connect the p^{th} node in X_i, x_{ip}, and the q^{th} node in X_j, x_{jq}, where $i \in \{1, \ldots, m\}$ and $j \in \{m+1, \ldots, m+k\}$, whenever the data object x_{ip} has a non-zero feature value of the feature represented by x_{jq}. The edge weight is set to be the feature value.

F-RankClass relates data objects with each other through features. Data objects of the same class usually have similar connections to feature nodes. Therefore, class information is more likely to propagate between similar data objects than between non-similar data objects. F-RankClass propagates class information between data objects based on feature similarities. The more similar two objects are, the more class information is propagated between.

To sum up, F-RankClass extends RankClass to a general framework, which can be applied to either a binary or multi-class classification problem of multimodal or unimodal data.

D. Building Heterogeneous Information Networks for F-RankClass from Multimodal Documents

There are two main components in multimodal documents: text and images. Given a multimodal document corpus, we denote by T the set of text nodes, I the set of image nodes,

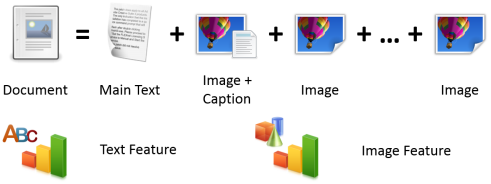


Fig. 2: A multimodal document consists of a piece of *main text* T, a set of *images* I, and possibly with some captions. Text features are extracted from the main text and captions. Image features are extracted only from images.

TF the set of text feature nodes, and IF the set of image feature nodes. Unigram text features [12] are extracted from both the main text and the captions of images. Dense SIFT Bag-of-features image features [13] are extracted from images. Note that some images have captions while others do not. Fig. 2 shows the legend of our figures.[1]

Text features are extracted by a bag-of-words model [12] [14], and image features are extracted by a dense SIFT bag-of-feature model [15] [13].

We consider two scenarios in the following.

1) Building Heterogeneous Information Networks from Unimodal Documents: We first extract features from the text-only documents and then connect a text node $t \in T$ to a text feature node $tf \in TF$ if t has a non-zero feature value tf. The weight of the edge is the feature value. A text-based information network is then built. In fact, the term-frequency matrix itself is the relation matrix. Fig. 3a illustrates the resulting network.

2) Building Heterogeneous Information Networks from Multimodal Documents: We first extract 1) text features from main text and image captions and 2) image features from images. Text nodes and image nodes are connected to their corresponding text feature and image feature nodes. We also connect nodes representing text and images that coexist in a document. Fig. 3b illustrates the the resulting network. If image captions are missing or misleading, a variation without image captions can be used.

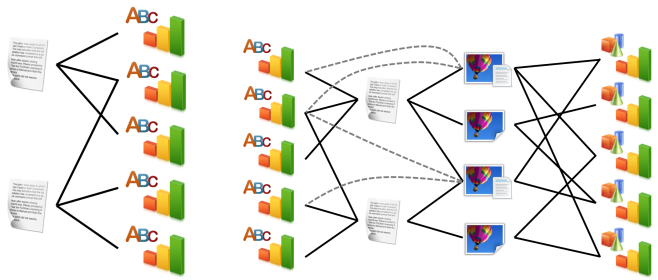
E. Comments on the Complexity of F-RankClass

The complexity of RankClass is $O(N_1K(|E| + |V|) + N_2K|V|)$, where N_1 is the number of iterations for computing the ranking distribution, N_2 is the number of iterations for the EM algorithm, K is the number of classes, $|E|$ and $|V|$ is the number of edges and nodes in the network, respectively.

F-RankClass introduces feature nodes and their connections with data object nodes. The extra number of edges, compared to RankClass, depends on the sparsity of features. Let at most $\hat{d}$ out of d features have non-zero values. F-RankClass introduces $O(\hat{d}|V|)$ extra edges and d extra nodes. Therefore, the complexity of F-RankClass is:

$$O(N_1K(|E| + \hat{d}|V| + |V| + d) + N_2K(|V| + d)). \quad (6)$$

[1]Icons designed by Tango, Harwen Zhang, Kyo Tux, Custom Icon Design, and Oxygen Team.



(a) (b)

Fig. 3: (a) An heterogeneous information network constructed from a text corpus, consisting of *text* nodes and *text feature* nodes. Edges are constructed from the term-frequency matrix. (b) An heterogeneous information network constructed from multimodal documents, consisting of *text* nodes, *image* nodes, *text feature* nodes, and *image feature* nodes. A caption is connected to its associated image and its related text feature nodes with feature values as edge weights. The edge weights between features and text/images are the feature values of the text/images.

IV. Experimental Results

A. Multimodal Document Dataset Collection

We experimented on 2008/9 Wikipedia Selection for Schools [11], which has around 5500 documents and 15 categories in total. Each document contains text and around 5-20 images and is manually assigned a category. We selected 3 datasets from 2008/9 Wikipedia for Schools, as shown in Table I. Datasets 1–3 represent an easier small dataset, a more challenging large dataset, and a multiclass classification scheme, respectively. We also take the text portion for unimodal text classification experiments.

TABLE I: Selected datasets from 2008/9 Wikipedia Selection for Schools.

Dataset No.	Category	No. of Documents	No. of Images
1	art	86	602
	information technology	86	354
2	history	836	6388
	science	1205	7226
3	business	141	1273
	language and literature	197	842
	math	263	1027
	religion	176	1029

B. Experimental Setup

We performed unimodal and multimodal classification experiments on the datasets. Unimodal experiments are summarized in Section IV-C. Multimodal experiments with different graph structures and parameter settings are summarized in Section IV-D. F-RankClass is compared with a baseline text-based SVM and a SVM-based meta-classifiers [5] on unimodal and multimodal tasks, respectively. Classification accuracy is reported and used as the evaluation criteria.

Model selection of both F-RankClass and SVM is done on a validation set. For F-RankClass, we pick the best model among iterations. For SVM, we pick the trade-off parameter, C, between training error and margin width.

TABLE II: Accuracy comparison between F-RankClass and linear SVM on textual data. Numbers in the table represent accuracy.

	Dataset 1	Dataset 2	Dataset 3
Linear-SVM	93.3%	89.2%	68.7%
F-RankClass	**100.0%**	**96.6%**	**96.0%**

All experiments on Wikipedia use the same training-test-validation split. Ten documents per class are set as the training data. We then take 20% of the remaining documents in each class as a validation set. The rest are then used as the test set.

C. Unimodal Classification by F-RankClass

We present experimental results of F-RankClass and compare with results of SVM on unimodal datasets. F-RankClass and linear SVMs are evaluated on Datasets 1-3. We pick the parameters of F-RankClass as follows. All λ_{ij} are set to 0.2 and all α_i are set to 0.1. Results on accuracy are summarized in Table II.

F-RankClass outperforms linear SVMs on all 3 datasets. On Dataset 1, F-RankClass gives 100% accuracy. On Dataset 3, F-RankClass yields 27.3% better accuracy than linear SVMs.

D. Multimodal Document Classification Based on F-RankClass

1) Study on Graph Structures: We compare results of F-RankClass on four different settings. The first setting is the same as in Subsection IV-C. Only text and text features are used to build networks. The second setting uses text, text features, image, and image features to build networks, but not captions; hence, images are not connected to text features. The third setting uses all information except text-image links. The fourth setting uses all information, as described in Subsection III-D. The four settings are summarized in Table III.

TABLE III: Different settings on building the graph from 2008/9 Wikipedia Selection for Schools.

Setting No.	Text	Image	Captions	Text-Image Links
1	Yes	No	No	No
2	Yes	Yes	No	Yes
3	Yes	Yes	Yes	No
4	Yes	Yes	Yes	Yes

We compare classification results of F-RankClass with text-based linear SVMs and image-based SVM with intersection kernels on Datasets 1-3. The SVM-based meta-classifier proposed by Chen *et al.* [5], which utilizes image information to aid text classification, is also implemented for comparison. Text and image classification results are summarized in Tables IV and V, respectively.

F-RankClass performs better than linear SVMs on all four settings. For text classification, Settings 1 and 2 give the highest accuracy. No setting works the best across all datasets for text and image classification. This is because the text-image relationships in each dataset have different natures. In some datasets, text, images, and captions in a document are all highly correlated, so more links in the graph enhance classification results. In some other datasets, text and images

TABLE IV: Text classification accuracy comparison between F-RankClass on different graph structures. Results from linear SVMs and the SVM-based meta-classifier proposed in [5] are also included as baselines. Numbers in the table represent accuracy.

	Dataset 1	Dataset 2	Dataset 3
1	**100.0**%	96.6%	**96.0**%
2	**100.0**%	**96.9**%	**96.0**%
3	**100.0**%	96.4%	95.0%
4	**100.0**%	94.0%	92.4%
SVM	93.3%	89.2%	68.7%
SVM-Meta	86.7%	86.2%	76.3%

TABLE V: Image classification accuracy comparison between F-RankClass on different graph structures. Results from linear SVMs are also included as a baseline. Numbers in the table represent accuracy.

	Dataset 1	Dataset 2	Dataset 3
2	**93.5**%	54.4%	30.7%
3	71.2%	**74.4**%	**56.1**%
4	79.9%	73.3%	48.4%
SVM	80.5%	59.8%	29.7%

or images and captions provide complementary information. Therefore, absence of links helps maintain correct labeling by not propagating conflicting information.

Although the image classification accuracy of Dataset 3 from linear SVMs is less than 30%, the SVM-based meta-classifier in [5] is still able to outperform linear SVMs in text classification on Dataset 3 because of the extra image information. This shows the importance of utilizing multimodal relationships in multimodal data mining. Note that the SVM-based meta-classification framework uses image information to aid text classification and classifies documents as a whole; therefore, the text-image combined classification accuracies are reported in Table IV.

2) Varying Cross-Type Influence λ: In RankClass, the parameter $\lambda_{ij}, i, j \leq m$ controls information flows between nodes of types i and j; thus affecting the ranking distribution $P(x_{ip}|X_i, k)$ as in Eq. 2. We performed experiments on Datasets 1-3 with seven different sets of parameters $\Lambda_1, \ldots, \Lambda_{4b}$, as summarized in Table VI. Parameter set Λ_1 is the default setting. Parameter sets $\Lambda_{2a}, \Lambda_{3a}, \Lambda_{4a}$ set all but one parameter to 0.2. Parameter sets $\Lambda_{2b}, \Lambda_{3b}, \Lambda_{4b}$ set all but one parameter to 0.4. In Λ_{2a} and Λ_{2b}, $\lambda_{1,2}$ (text-image linkage) is tuned down to 0.1. In Λ_{3a} and Λ_{3b}, $\lambda_{2,3}$ (image-caption linkage) is tuned down to 0.1. In Λ_{4a} and Λ_{4b}, $\lambda_{2,4}$ (image-IF linkage) is tuned down 0.1. Note that we did not tune λ_{13} (text-TF linkage) because text features are already shown to be highly related with main text and classes of documents.

We compare classification accuracy for $\Lambda_1, \ldots, \Lambda_{4b}$ on Datasets 1-3. The text and image classification results are summarized in Table VII and Table VIII, respectively.

Some λ_{ij} affects text and image classification accuracy differently. For text classification, parameter sets Λ_{2a} and Λ_{2b} give the best results. This can be explained by the fact that images appear more randomly than text in a document. Some images act as complementary or supplementary information in a document and are not so coherent with the main contents. As a result, weakened text-image links improve text classification

TABLE VI: Experimental settings of λ_{ij} in Section IV-D2.

Parameter Set	λ_{12}	λ_{13}	λ_{23}	λ_{24}	Semantic Meaning
Λ_1	0.2	0.2	0.2	0.2	Baseline
Λ_{2a}	0.1	0.2	0.2	0.2	Weakened text-image links
Λ_{2b}	0.1	0.4	0.4	0.4	Weakened text-image links
Λ_{3a}	0.2	0.2	0.1	0.2	Weakened image-caption links
Λ_{3b}	0.4	0.4	0.1	0.4	Weakened image-caption links
Λ_{4a}	0.2	0.2	0.2	0.1	Weakened image-IF links
Λ_{4b}	0.4	0.4	0.4	0.1	Weakened image-IF links

TABLE VII: Text classification accuracy comparison between F-RankClass on different λ settings. Numbers in the table represent accuracy.

	Dataset 1	Dataset 2	Dataset 3
1	**100.0**%	94.0%	92.4%
2a	**100.0**%	**96.0**%	94.0%
2b	98.3%	96.4%	95.0%
3a	**100.0**%	92.1%	93.2%
3b	**100.0**%	89.3%	92.6%
4a	**100.0**%	95.6%	93.4%
4b	**100.0**%	95.4%	**94.2**%

accuracy.

On the other hand, there is no parameter set that performs the best across all datasets in terms of image classification. This is consistent with our observation on image classification with different graph structures: the relationships between text, images, and captions are dataset-dependent. Not only the existence of a type of edges, but also the amount of information flowing through the edges, plays an important role in the classification process. For example, parameter sets Λ_{3a} and Λ_{3b} represent weakened image-caption links, with all other λ being set to either 0.2 or 0.4. On Dataset 3, Λ_{3b} performs the best, while Λ_{3a} does not work particularly well. For another example, while Graph Structure Setting 2, which excludes captions, works the best on Dataset 1, parameter sets Λ_{4a} and Λ_{4b}, which weaken image-IF but not image-caption links, give the best accuracy on Dataset 1. This indicates that λ_{ij} needs to be carefully selected to control the information flow for optimal classification results.

E. Model Selection of F-RankClass

As with any other classification algorithms, F-RankClass suffers from overfitting as the number of iterations increases. Figure 4 shows text classification accuracies on the test and validation set of Dataset 3 throughout 25 iterations, which represent a typical trend of classification accuracy *vs.* iterations in various datasets. Accuracy often reaches the peak between

TABLE VIII: Image classification accuracy comparison between F-RankClass on different λ settings. Numbers in the table represent accuracy.

	Dataset 1	Dataset 2	Dataset 3
1	79.9%	73.3%	48.4%
2a	75.9%	75.2%	57.5%
2b	75.5%	**76.9**%	58.9%
3a	76.1%	73.9%	52.7%
3b	75.5%	75.1%	**62.0**%
4a	80.1%	70.6%	46.4%
4b	**86.9**%	68.0%	43.3%

iterations 2-6, and then drops significantly due to overfitting. When overfitting occurs, the structure of the underlying heterogeneous information network is dominated by the training data. Edge weights that represent feature values are mostly close to zero because of the update equation, Eq. 3. As a result, accuracy does not recover in the following iterations.

The accuracy of validation sets follow a trend similar to the accuracy of testing sets. This justifies our model selection criteria. Although the highest accuracies of testing and validation sets may not occur in exactly the same iteration, one is still able to select a reasonably good model.

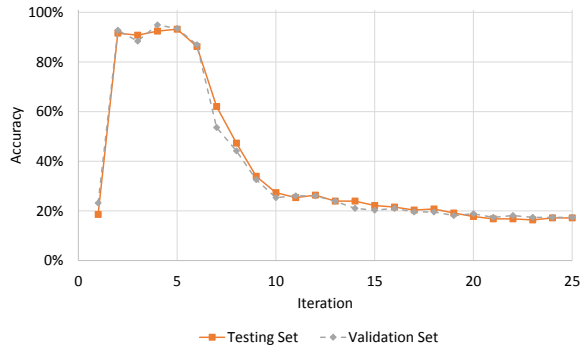


Fig. 4: Text classification accuracy vs. F-RankClass iterations on Dataset 3. There exists high consistency in the accuracy of the testing set and the accuracy of the validation set.

F. Discussion on Joint Text-Image Relationships

The experiments on F-RankClass provide insights on how text and images in multimodal documents interact with each other. Several observations are made from the experimental results presented above.

First of all, text-image relationships vary significantly from dataset to dataset. For example, Table IV shows that the absence of edges between text and images worsens image classification accuracy on Dataset 1, but it enhances the accuracy on Datasets 2 and 3. On the contrary, the absence of captions undermines image classification accuracy on Datasets 2 and 3, while it helps improve accuracy on Dataset 1.

Secondly, edges between text and images plays an important role in the information network. In Tables IV and V, it is shown that a graph without direct text-image edges in general does not give good accuracy for both text and image classifications. Meanwhile, text classification accuracy improves when direct text-image linkage is present but *weakened*, as shown in Table VII. This indicates the importance to control the amount of information flowing between text and images.

A similar phenomenon is observed on captions. Image classification accuracy often improves when direct image-caption edges is present but *weakened*, as shown in Table VIII. Meanwhile, the total absence of captions usually worsens the overall image classification accuracy, as shown in Table V.

Finally, text and image classification may work the best in different parameter settings. In Tables VII and VIII, we show that weakened text-image edges helps improve text classification accuracy, while this setting does not always improve image classification accuracy. This implies text and image classification may need to be done in separate settings to obtain optimal results.

V. Conclusion

To summarize, we propose a novel heterogeneous information network mining framework, *F-RankClass*, which is designed to work with both multimodal and unimodal data. We performed text classification and multimodal document classification experiments and compared with previous work. F-RankClass is superior to its counterparts in text and multimodal document classification. The experimental results also provide insights on joint text-image relationships. Furthermore, F-RankClass is not limited to the applications we discussed. Most classification problems, either binary or multiclass, unimodal or multimodal, fit into the F-RankClass framework.

References

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[2] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining Text Data*. Springer, 2012, pp. 163–222.

[3] N. Chen, "A survey of indexing and retrieval of multimodal documents: Text and images," School of Computing, Queen's University, Tech. Rep. 2006-505, 2006.

[4] H. Shatkay, N. Chen, and D. Blostein, "Integrating image data into biomedical text categorization," *Bioinformatics*, vol. 22, no. 14, pp. e446–e453, 2006.

[5] S. D. Chen, V. Monga, and P. Moulin, "Meta-classifiers for multimodal document classification," in *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*. IEEE, 2009, pp. 1–6.

[6] M. Maree, S. M. Alhashmi, M. Belkhatir, and A. Hawit, "Automatic construction of a domain-independent knowledge base from heterogeneous data sources," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*. IEEE, 2012, pp. 1483–1488.

[7] T. S. Lee, S. Fidler, A. Levinshtein, and S. Dickinson, "Learning categorical shape from captioned images," in *Computer and Robot Vision (CRV), 2012 Ninth Conference on*. IEEE, 2012, pp. 228–235.

[8] M. Ji, J. Han, and M. Danilevsky, "Ranking-based classification of heterogeneous information networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1298–1306.

[9] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma, "Object-level ranking: bringing order to web objects," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 567–574.

[10] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 797–806.

[11] Wikipedia Foundation. (2008) 2008/9 Wikipedia selection for schools. [Online]. Available: http://http://schools-wikipedia.org/

[12] T. Joachims, *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[14] D. Zeimpekis, E. M. Kontopoulou, and E. Gallopoulos. (2008, Dec.) Text to matrix generator. [Online]. Available: http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG/

[15] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.