Large-Scale Cross-Language Web Page Classification via Dual Knowledge Transfer Using Fast Nonnegative Matrix Trifactorization

HUA WANG, Colorado School of Mines
FEIPING NIE and HENG HUANG, University of Texas at Arlington

With the rapid growth of modern technologies, Internet has reached almost every corner of the world. As a result, it becomes more and more important to manage and mine information contained in Web pages in different languages. Traditional supervised learning methods usually require a large amount of training data to obtain accurate and robust classification models. However, labeled Web pages did not increase as fast as the growth of Internet. The lack of sufficient training Web pages in many languages, especially for those in uncommonly used languages, makes it a challenge for traditional classification algorithms to achieve satisfactory performance. To address this, we observe that Web pages for a same topic from different languages usually share some common semantic patterns, though in different representation forms. In addition, we also observe that the associations between word clusters and Web page classes are another type of reliable carriers to transfer knowledge across languages. With these recognitions, in this article we propose a novel joint nonnegative matrix trifactorization (NMTF) based Dual Knowledge Transfer (DKT) approach for cross-language Web page classification. Our approach transfers knowledge from the auxiliary language, in which abundant labeled Web pages are available, to the target languages, in which we want to classify Web pages, through two different paths: word cluster approximation and the associations between word clusters and Web page classes. With the reinforcement between these two different knowledge transfer paths, our approach can achieve better classification accuracy. In order to deal with the large-scale real world data, we further develop the proposed DKT approach by constraining the factor matrices of NMTF to be cluster indicator matrices. Due to the nature of cluster indicator matrices, we can decouple the proposed optimization objective and the resulted subproblems are of much smaller sizes involving much less matrix multiplications, which make our new approach much more computationally efficient. We evaluate the proposed approach in extensive experiments using a real world cross-language Web page data set. Promising results have demonstrated the effectiveness of our approach that are consistent with our theoretical analyses.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Cross-language classification, nonnegative matrix trifactorization, knowledge transfer, cluster indicator matrix, large-scale data

ACM Reference Format:
Hua Wang, Feiping Nie, and Heng Huang. 2015. Large-scale cross-language web page classification via Dual Knowledge Transfer using fast nonnegative matrix trifactorization. ACM Trans. Knowl. Discov. Data 10, 1, Article 1 (July 2015), 29 pages.
DOI: http://dx.doi.org/10.1145/2710021

This work was partially supported by NSF-IIS 1117965, NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1423591.
Authors' addresses: H. Wang, Department of Electrical Engineering and Computer Science, Colorado School of Mines; email: huawang@mines.edu; F. Nie and H. Huang (corresponding author), Department of Computer Science, University of Texas at Arlington; emails: feipingnie@gmail.com, heng@uta.edu.

DOI: http://dx.doi.org/10.1145/2710021

1. INTRODUCTION

With the rocketing growth of Internet in recent years, an ever-increasing number of Web pages are now available in many different languages. As of July 2013, over 145.7 million web sites are actively in operation[1], with billions of Web pages created in almost all human languages. As a result, cross-language information retrieval (IR) has become unprecedentedly important for organizing and mining information stored in Web pages in various languages.

A potential problem in categorizing Web pages, especially for those written in uncommonly used languages, is lacking sufficient labeled training data. This prevents one from training effective classification models, which usually require a large amount of labeled data. Statistically speaking, the more labeled training data one can use, the more accurate and robust classification models one can build. Fortunately, due to many reasons, there exist a lot of labeled Web pages in several most commonly used languages, such as English. Examples of these resources include Reuters-21578[2], 20-Newgroups[3], Open Document Project[4], and many others. It is thus useful and intriguing to make use of these labeled Web pages in one language, called as *auxiliary language*, to help classify Web pages in another language, called as *target language*. This problem is called as *cross-language Web page classification* [Ling et al. 2008], which has aroused a lot of interests in recent researches. In this article, we explore this new, yet important, knowledge discovery problem by proposing a novel joint NMTF based DKT approach.

1.1. Challenges in Cross-Language Web Page Classification

One of the most widely used strategy in cross-language Web page (text) classification is using language translation [Ling et al. 2008; Olsson et al. 2005; Prettenhofer and Stein 2010; Ramírez-de-la Rosa et al. 2010; Shi et al. 2010; Wan 2009; Wu and Lu 2008]. One can either translate the test data of a problem into the auxiliary language, or translate the training data into the target language. Then one can train and classify the resulted data in one single language. Although this straightforward method may be feasible, it suffers from a number of critical problems that impede its practical use [Ling et al. 2008; Ramírez-de-la Rosa et al. 2010; Shi et al. 2010]. Thus we first examine the challenges in cross-language Web page classification and seek opportunities to overcome them, which motivate our approach.

1.1.1. Cultural Discrepancy. The first difficulty in cross-language Web page classification is caused by cultural discrepancies, which heavily impacts the classification performance in spite of a perfect translation [Ling et al. 2008; Ramírez-de-la Rosa et al. 2010]. Given that a language is the way to express a cultural and socially homogenous community, Web pages from a same category but different languages may concern very different topics. For example, let us consider Web pages that report sports news in France (in French) and in USA (in English). While the former typically pay more attention to soccer, rugby and cricket, the latter are more interested in basketball and American football. From machine learning perspective of view, this is the situation in which the training data and test data are drawn from different distributions, which makes it a challenge for traditional supervised and semi-supervised classification algorithms to achieve satisfactory Web pages classification performance.

[1] http://www.domaintools.com/internet-statistics/.
[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/.
[3] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html.
[4] http://www.dmoz.org/.

W1: An Algorithm for Hyperlink Clustering
W2: Algorithms for Webpage Classification
W3: Texture Clustering Algorithms
W4: An Algorithm for Illumination Classification

(a) A synthetic data set of 4 Web pages in target language.

	Clustering	Classification	Illumination	Texture	Webpage	Hyperlink
W1	1	0	0	0	0	1
W2	0	1	0	0	1	0
W3	1	0	0	1	0	0
W4	0	1	1	0	0	0

(b) The original word-document co-occurrence representation of the data set.

		Learning		Graphics		Web	
		Clustering	Classification	Illumination	Texture	Webpage	Hyperlink
IR	W1	1		0		1	
IR	W2	1		0		1	
Vision	W3	1		1		0	
Vision	W4	1		1		0	

(c) The transformed representation of the data set by incorporating knowledge learned from auxiliary language, which leads to meaningful clustering results.

Fig. 1. An illustrative example to demonstrate the usefulness of leveraging the knowledge learned from auxiliary language when clustering Web pages in target language.

To overcome this problem, instead of simply combining the data in different languages, we consider to transfer labeled information contained in the Web pages in the auxiliary language to those in the target language [Pan and Yang 2009]. Our approach is based on the observation that Web pages in different languages from a same category often share some same semantic patterns, although they are in different representation forms, for example, the respective basic linguistic units in French and English [Ling et al. 2008]. Therefore, we may abstract the prior knowledge in the auxiliary language into certain semantic patterns, then make use of them to help classify Web pages in the target language. To transfer knowledge across languages, the most natural carrier is the basic linguistic representation unit—words. As shown in Figure 1, we use an example to illustrate the usefulness of knowledge transfer by word (KNW) clusters in Web page classification tasks.

Given a data set with four Web pages ($W1$, $W2$, $W3$, and $W4$) as shown in Figure 1(a), we represent them as a word-document matrix as shown in Figure 1(b). Because in practice we usually do not have labels for Web pages in the target language, we cluster them (the rows of the data matrix) based on the cosine similarity, which results in two clusters, ($W1$) and ($W3$) as a cluster and ($W2$) and ($W4$) as a cluster. This result, however, is not meaningful. If we use the learned knowledge from the auxiliary language to guide this clustering process, we can transform the data matrix by using the three semantic "hyper-features" as in Figure 1(c). That is, "clustering" and "classification" belong to "learning", "illumination" and "texture" belong to "graphics", and "webpage" and "hyperlink" belong to "Web". Clustering on this new transformed data matrix, we obtain ($W1$) and ($W2$) as a cluster and ($W3$) and ($W4$) as a cluster, which is a very meaningful result. Obviously, the topic of the former cluster is "IR", while that of the latter is "Vision". We will further discuss this example and give more theoretical analysis on it later in Section 3.2.

With earlier recognition, the first path in our new approach to transfer knowledge from the auxiliary language to the target one is by word patterns, that is, feature clusters, learned from the auxiliary language. This path is schematically shown by the red lines in Figure 2.

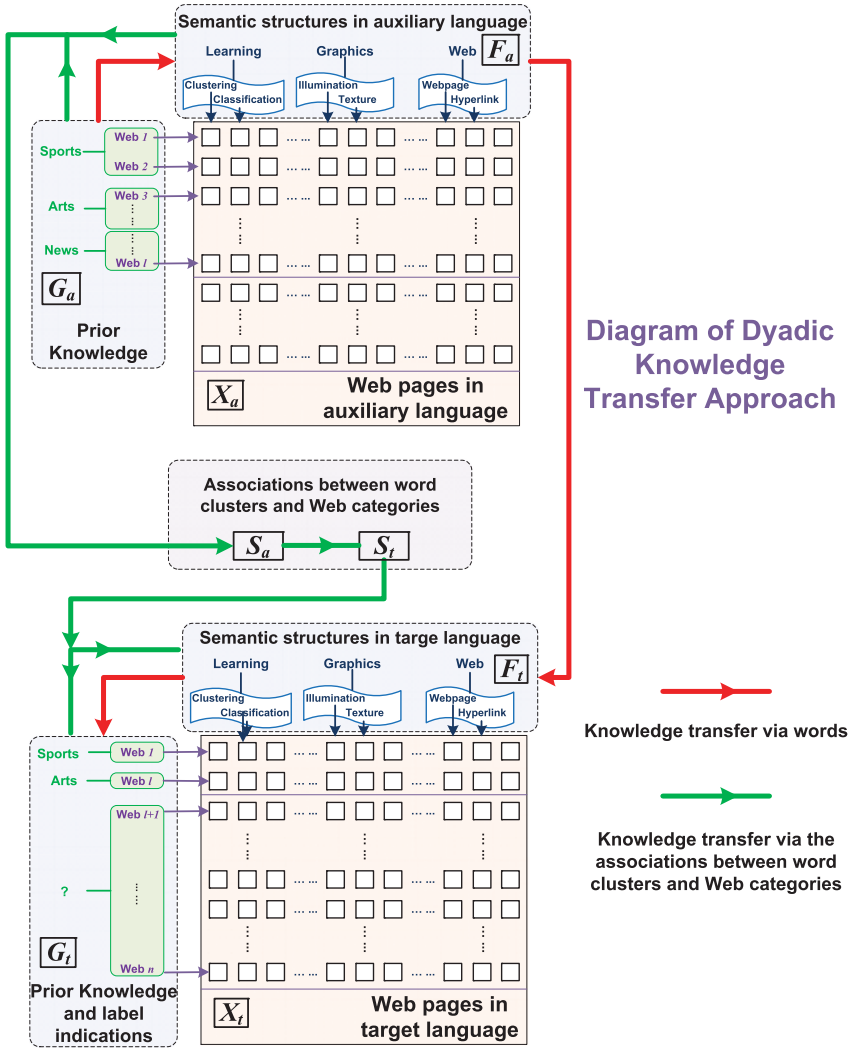


Fig. 2. Diagram of the proposed Dual Knowledge Transfer (DKT) approach using joint NMTF. We transfer knowledge from auxiliary language to target language through two ways: *words* (via $\mathbf{F}^a$ and $\mathbf{F}^t$) and *the associations between word clusters and Web page categories* (via $\mathbf{S}^a$ and $\mathbf{S}^t$).

1.1.2. Translation Ambiguity. During language translation, the ambiguities introduced by dictionaries are another challenge in cross-language Web page classification. For example [Zhuang et al. 2010], the word "阅读材料 (reading materials)" in Chinese Web pages could be reasonably translated as "textbooks", "required reading list", "reference", to name a few. Since the linguistic habits in expressing a concept are different in different languages, the phrases for a same concept may have different probabilities to appear in different languages. Therefore, transferring knowledge by the raw words sometimes are not reliable. In contrast, the concept behind the phrases may have the same effect to indicate the semantic class labels of the Web pages in different languages. In the same example, a Web page is more probable to be course-related if it contains the concept of "reading materials", no matter which specific key word (*i.e.*, "textbooks", "required reading list" or "reference") is being used. In other words, only

the concept behind raw words are stable in indicating taxonomy, and the associations between word clusters and Web page categories is independent of languages [Zhuang et al. 2010]. Therefore, we make use of it as the second bridge to transfer knowledge across different languages, which is illustrated by the green paths in Figure 2.

1.1.3. Data Diversity. One more challenge in cross-language Web page classification is the data diversity. As illustrated in Figure 2, although we may have a lot of training Web pages in one language, usually not all of them are fully labeled. Similarly, even the labeled resources in the target language are scarce, we may still have a small number of Web pages in this language, which are labeled by limited human efforts. Most importantly, even we have sufficiently many labeled data in the target language, due to the differences of culture and social focus, they might not cover all the Web page categories. Consider that, for example [Olsson et al. 2005], the English speakers tend to contribute more to some topics than their Czech counterparts (*e.g.*, to discuss "London" more than "Prague"). As a result, with data only in English we may expect to do poorly at identifying topics like "Prague". In contrast, Czech speakers often talk about "Prague". Thus, we may expect to improve on detecting topic "Prague" in English Web pages by leveraging Czech data. In conclusion, we cannot rigidly assume the Web pages in the auxiliary language are always labeled while the Web pages in the target language are not labeled at all. Namely, model flexibility must be addressed to handle real world cross-language Web page classification problems.

1.1.4. Algorithm Scalability. The last, but not least, challenge in Web page classification is the scalability of the algorithm. Because the hardware costs in contemporary systems keep to decrease in a fast speed, input data coming from real world applications are usually of large sizes, especially for those related to Internet. Therefore, it is critical to take into account the capability to deal with large-scale data, when one devises new Web page classification methods for practical use.

1.2. Our Model

Taking into account the four major challenges in cross-language Web page classification as detailed earlier, we propose a novel joint NMTF framework to abstract the prior knowledge contained in Web pages in the auxiliary language, including both labeling information by human efforts and latent language structures. That is, our new framework first represents the knowledge contained in the auxiliary language in two forms by the two factor matrices $\mathbf{F}^a$ and $\mathbf{S}$ of NMTF respectively. Then, it transfers the abstracted knowledge to the target language to guide the classification therein. The whole idea is schematically summarized in Figure 2. Because we employ two separate way to transfer knowledge, we call our new framework as the *DKT* approach.

Same as other existing nonnegative matrix factorization (NMF) and NMTF based clustering and classification methods [Chen et al. 2009; Ding et al. 2005, 2006, 2010; Gu and Zhou 2009a; Gu et al. 2010; Li et al. 2010; Wang et al. 2008], when traditional nonnegative constraints are used, the algorithm to solve the objective of the proposed DKT method involves intensive matrix multiplications, which make it computationally inefficient. In order to deal with large-scale real world data, following the same idea in our previous work [Wang et al. 2011a], we also present a fast version of the proposed DKT approach. To be more specific, instead constraining the factor matrices of NMTF to be nonnegative, we constrain them to be cluster indicator matrices, a special type of nonnegative matrices. As a result, the new optimization problem can be decoupled, which results in subproblems of much smaller sizes requiring much less matrix multiplications, such that our approach scales well to large-scale input data. Moreover, the resulted factor matrices can directly assign class labels to Web pages due to the nature of indicator matrices. In contrast, existing NMF based methods have to require an

extra post-processing step to extract cluster structures from the factor matrices, which often leads to non-unique clustering results.

We summarize our contributions as following.

—To address the cross-language Web page classification problem, we observe the two possible ways to transfer knowledge across languages: the natural way by *word clusters* and the reliable way by *associations between word clusters and Web page categories*. We propose a joint NMTF based DKT approach to make use of these two ways, thus improve the classification performance.
—Through the general framework of the proposed approach, we consider a variety of conditions in cross-language Web page classification. Regardless the amount of labeled training data and locations in which they are, either in auxiliary language or target language, or the both, our approach is always able to take advantage of the available information.
—In order to deal with large-scale real world data, a fast version of the proposed approach with high computational efficiency is presented, which constrains the factor matrices of NMTF to be cluster indicator matrices, instead of nonnegative as in existing methods.
—Extensive experiments on real world data sets demonstrate promising results that validate our approach.

This work expands on our previous conference publication [Wang et al. 2011b] and differs from it in the following two important aspects. First, when transferring the prior labeling knowledge in the auxiliary language to the target language, instead of forcing the associations between word patterns and semantic classes to be identical in both languages, we approximate the associations in the target language to that in the auxiliary language. As a result, our new objective is more general and flexible, which thereby could more accurately model the real world applications and alleviate the impact caused by the semantic class difference between the two languages. Second, in this manuscript, we further develop the proposed DKT approach and present a fast version that is much more computationally efficient and scales well to large scale real world Web page data. Constraining the factor matrices of NMTF to be cluster indicator matrices thereby introducing the fast DKT approach is the major difference of this work compared to our previous conference paper [Wang et al. 2011b]. In addition, we also provide additional experimental results to demonstrate the effectiveness and efficiency of both the proposed DKT approach and its fast version in cross-language Web page classification.

1.3. Article Organization

The rest of this article is organized as following. We first introduce some backgrounds and briefly review co-clustering via NMTF in Section 2, by which we motivate our approach. Then in Section 3 we systematically develop the objective of the proposed DKT approach. In order to improve the computational efficiency of the proposed DKT approach to deal with large-scale real world data, we further introduce a fast version of the proposed DKT approach in Section 4, which is one of the major contribution of this article compared to our previous conference paper [Wang et al. 2011b]. The connections of the proposed approach and related existing works are examined in Section 5. Finally, we report experimental results to evaluate a variety aspects of the proposed approaches in Section 6 and conclude our work in Section 7.

2. BACKGROUNDS

In this section, we first introduce some notations, which will be frequently used in this article. Then we briefly review NMTF for co-clustering and reveal how it transfers

knowledge between data and features within the same data set, from which we will develop the proposed approach in next section.

2.1. Notations

Throughout this article, we denote matrices as bold uppercase characters and vectors as bold lowercase characters. For the matrix $\mathbf{M}$, we denote its entry at the ith row and jth column as $\mathbf{M}_{(ij)}$. The ith row and jth column of $\mathbf{M}$ are denoted as $\mathbf{m}_{i\cdot}$ and $\mathbf{m}_{\cdot j}$ respectively. We denote the Frobenius norm and the trace of the matrix $\mathbf{M}$ as $\|\mathbf{M}\|$ and $\mathbf{tr}(\mathbf{M})$ respectively.

The real number set is denoted as $\Re$ and the nonnegative real number set is denoted as $\Re_+$.

In particular, in this article we use Ψ to denote the set of *cluster indicator matrices*. A cluster indicator matrix $\mathbf{G} \in \Psi^{n\times c}$ is a special type of nonnegative matrix satisfying the following requirement: given a row $\mathbf{g}_{i\cdot}$ $(1 \le i \le n)$ of $\mathbf{G}$, all its entries are equal to 0 except for one and only one entry equal to 1, which indicates the cluster membership of the ith data point. To be more specific, for the ith row of $\mathbf{G}$, the following two conditions are satisfied: $\mathbf{g}_{i\cdot} \in \{0, 1\}^c$ and $\sum_j \mathbf{g}_{i\cdot}(j) = 1$.

2.2. Review of Co-Clustering via NMTF

Traditional clustering methods focus on one-side clustering, that is, clustering the data side based on the similarities along the feature side. In co-clustering problems [Ding et al. 2006; Wang et al. 2011c, 2011d], we cluster data points based on the distributions of features, meanwhile we cluster features based on the distributions of data points. Formally, given a data set $\mathcal{X} = \{\mathbf{x}_{\cdot i} \in \Re^d\}_{i=1}^n$, we write $\mathbf{X} = [\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot n}] = [\mathbf{x}_{1\cdot}^T, \ldots, \mathbf{x}_{d\cdot}^T]^T$. The goal of co-clustering is to group the data points $\{\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot n}\}$ into k_1 clusters $\{\mathcal{C}_j\}_{j=1}^{k_1}$, and simultaneously group the features $\{\mathbf{x}_{1\cdot}, \ldots, \mathbf{x}_{d\cdot}\}$ into k_2 clusters $\{\mathcal{W}_j\}_{j=1}^{k_2}$.

K-means clustering is a standard clustering method in statistical learning, which partitions the input data points into k disjoint clusters by minimizing the following objective [Ding and He 2004]:

$$J_{K\text{-means}} = \sum_{j=1}^{k} \sum_{\mathbf{x}_{\cdot i} \in \mathcal{C}_j} \|\mathbf{x}_{\cdot i} - \mathbf{c}_j\|^2 = \sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} \|\mathbf{x}_{\cdot i} - \mathbf{c}_j\|^2, \quad s.t. \quad \mathbf{G} \in \Psi^{n\times k}, \tag{1}$$

where $\mathbf{c}_j$ is the centroid of the jth cluster of the input data $\mathbf{X}$. Because $\mathbf{G} \in \Psi$ is a cluster indicator matrix, minimizing $J_{K\text{-means}}$ is a combinatorial optimization problem, which is hard to resolve in general. To tackle this, the problem of minimizing $J_{K\text{-means}}$ in Equation (1) can be relaxed to maximize the following objective [Ding and He 2004; Zha et al. 2001]:

$$J'_{K\text{-means}} = \mathbf{tr}(\mathbf{G}^T\mathbf{X}^T\mathbf{X}\mathbf{G}), \quad s.t. \quad \mathbf{G}^T\mathbf{G} = I. \tag{2}$$

Note that, in Equation (2) $\mathbf{G}$ in $J'_{K\text{-means}}$ is no longer an indicator matrix, but an arbitrary orthonormal matrix. The orthonormality of $\mathbf{G}$ guarantees the uniqueness of the solution of Equation (2), and leads to the clustering interpretation of $\mathbf{G}$ [Ding and He 2004; Zha et al. 2001].

Recently, Ding et al. [2005] explored the relationship between the relaxed objective of K-means clustering in Equation (2) and NMF, and proposed to use NMTF to simultaneously cluster both the features and the data points of an input data set, which is called as *co-clustering* of the input data set. The original NMF [Lee and Seung 1999, 2001] aims to find two nonnegative matrices whose product can well approximate the original nonnegative data matrix $\mathbf{X} \in \Re_+^{p\times n}$, that is, $\mathbf{X} \approx \mathbf{F}\mathbf{G}^T$, where $\mathbf{F} \in \Re_+^{d\times k}$ and

$\mathbf{G} \in \Re_+^{n\times k}$. The columns of $\mathbf{X}$ are data points and the rows of $\mathbf{X}$ are features (observations). An appropriate objective of NMF is to minimize the following objective [Lee and Seung 2001]:

$$J_{\text{NMF}} = \|\mathbf{X} - \mathbf{F}\mathbf{G}^T\|^2, \quad s.t. \quad \mathbf{F} \geq 0, \mathbf{G} \geq 0. \tag{3}$$

According to Ding et al. [2005], NMF defined in Equation (3) corresponds to simultaneous K-means clustering of the rows (features) and columns (data points) of $\mathbf{X}$, where $\mathbf{F}$ can be considered as the clustering indications for features and $\mathbf{G}$ can be considered as the clustering indications for data points. Because both $\mathbf{F}$ and $\mathbf{G}$ are relaxed cluster indications, that is, continuous real-valued matrices, they are called as *soft labels* [Ding et al. 2005]. Because co-clustering on the both sides of an input data matrix takes advantage of the interrelatedness between the data points and features, NMF based co-clustering methods usually report superior performance [Ding et al. 2005, 2006].

Because two-factor NMF in Equation (3) is restrictive, which often gives a rather poor low-rank matrix approximation, one may introduce one additional factor matrix $\mathbf{S} \in \Re_+^{k_1\times k_2}$ to absorb the different scales of $\mathbf{X}$, $\mathbf{F}$, and $\mathbf{G}$, which leads to NMTF [Ding et al. 2006] minimizing the following objective:

$$J_{\text{NMTF}} = \|\mathbf{X} - \mathbf{F}\mathbf{S}\mathbf{G}^T\|^2 \quad s.t. \quad \mathbf{F} \geq 0, \mathbf{S} \geq 0, \mathbf{G} \geq 0, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{G}^T\mathbf{G} = \mathbf{I}, \tag{4}$$

where $\mathbf{F} \in \Re_+^{d\times k_1}$ and $\mathbf{G} \in \Re_+^{n\times k_2}$. $\mathbf{S}$ provides increased degrees of freedom such that the low-rank matrix representation remains accurate while $\mathbf{F}$ gives row clusters and $\mathbf{G}$ gives column clusters. Most importantly, $\mathbf{S}$ is a condensed view of $\mathbf{X}$ [Li et al. 2009] and represents the associations between word clusters and Web page clusters [Zhuang et al. 2010]. Note that here we impose the orthogonal constraints on the both side factor matrices, such that the solution of the problem is unique [Ding et al. 2006].

2.3. Motivations of Using NMTF in Cross-Language Web Page Classifications

In the context of Web page classification, the input data set is described by $\mathbf{X}$ for a set of Web pages in a given language. Due to their roles in co-clustering as introduced earlier, the factor matrices $\mathbf{F}$ and $\mathbf{S}$ of the NMTF in Equation (4) abstract the two types of information contained in $\mathbf{X}$: the former characterizes the feature clustering patterns of the input data, which are the unsupervised knowledge of $\mathbf{X}$ due to the intrinsic linguistic structures of the language; while the latter encodes the associations between feature clusters and semantic classes, which are the supervised knowledge due to human labeling efforts. Obviously, these two types of transformed knowledge are exactly what we expect to transfer across languages as outlined in Section 1. However, both Equations (3) and (4) are designed for one single data set, while in cross-language Web page classifications we have two separate data sets, one in the auxiliary language and the other in the target language. Thus, in next section we further develop NMTF in Equation (4) and propose a novel joint NMTF framework to transfer knowledge across languages to address the challenges in cross-language Web page classification to achieve improved classification performance.

3. JOINT NMTF BASED DUAL KNOWLEDGE TRANSFER

In this section, we develop a novel joint NMTF based DKT approach for cross-language Web page classification, which transfers knowledge from the data in the auxiliary language to those in the target language by two different paths: (1) word cluster approximation and (2) the associations between word clusters and Web page classes. An iterative algorithm to solve the proposed objective is also presented.

3.1. Problem Formalization of Cross-Language Web Page Classification

For a cross-language Web page classification problem, we have two Web page data sets, one in the auxiliary language $\mathbf{X}^a = [\mathbf{x}^a_{\cdot 1}, \ldots, \mathbf{x}^a_{\cdot n^a}] \in \Re_+^{d \times n^a}$ and the other in the target language $\mathbf{X}^t = [\mathbf{x}^t_{\cdot 1}, \ldots, \mathbf{x}^t_{\cdot n^t}] \in \Re_+^{d \times n^t}$, where $\mathbf{x}^a_{\cdot i}$ represents the ith Web page in the auxiliary language and $\mathbf{x}^t_{\cdot j}$ represents the jth Web page in the target language. Thus $\mathbf{X}^a$ and $\mathbf{X}^t$ can be seen as the document-word co-occurrence matrices of the auxiliary data and the target data respectively, or their *tf–idf* normalized counterparts. We assume that the both data sets are using a same vocabulary with d keywords: if the vocabularies differ, we may simply pad zeros in the feature vectors and re-express them under the same unified vocabulary so that the indices of the feature vectors from the both data sets correspond to the same word. Let $\mathbf{V} \in \Re^{d \times d}$ be a diagonal matrix with $\mathbf{V}_{(ii)} = 1$ if the ith word occurs in the both data sets, and $\mathbf{V}_{(ii)} = 0$ otherwise.

Typically a large amount of Web pages in the auxiliary language are manually labeled, which can be described by an indicator matrix $\mathbf{Y}^a \in \Re^{n^a \times k_2}$ such that $\mathbf{Y}^a_{(ik)} = 1$ if $\mathbf{x}^a_{\cdot i}$ belongs to the kth class, and $\mathbf{Y}^a_{(ik)} = 0$ otherwise. In addition, a diagonal matrix $\mathbf{C}^a \in \Re^{n^a \times n^a}$ is used, whose entry $\mathbf{C}^a_{(ii)} = 1$ if Web page $\mathbf{x}^a_{\cdot i}$ is labeled by the ith row of $\mathbf{Y}^a$, and $\mathbf{C}^a_{(ii)} = 0$ otherwise. Note that if $\mathbf{C} = \mathbf{I}$, all the Web pages in auxiliary language are completely labeled and specified by $\mathbf{Y}^a$. Sometimes, though not always, we may also have a limited number of labeled Web pages in the target language. We similarly describe them using $\mathbf{Y}^t \in \Re^{n^t \times k_2}$ such that $\mathbf{Y}^t_{(ik)} = 1$ if $\mathbf{x}^t_{\cdot i}$ belongs to the kth class, and $\mathbf{Y}^t_{(ik)} = 0$ otherwise. Again, $\mathbf{C}^t \in \Re^{n^t \times n^t}$ is a diagonal matrix whose entry $\mathbf{C}^t_{(ii)} = 1$ if Web page $\mathbf{x}^t_{\cdot i}$ is labeled by the ith row of $\mathbf{Y}^t$, and $\mathbf{C}^t_{(ii)} = 0$ otherwise. When the labels for all the Web pages in the target language are not available, we set $\mathbf{C}^t = 0^{n^t \times n^t}$, which is a zero matrix. Our goal is to predict labels for the unlabeled Web pages in the target data set.

In general, although the two sets of classes interested by the two data sets in the two languages overlap for most of the classes, they may differ. For the latter case, we pad zero columns to $\mathbf{Y}^a$ or $\mathbf{Y}^t$, or the both, such that the column indices of the both matrices correspond to the same classes. We encode the difference between the two sets of classes by two matrices, one for the auxiliary data and the other for the target data: $\mathbf{Q}^a \in \Re^{k_2 \times k_2}$ is a diagonal matrix with $\mathbf{Q}^a{}_{(ii)} = 1$ if the ith class comes from the source data set, and $\mathbf{Q}^a{}_{(ii)} = 0$ otherwise; and $\mathbf{Q}^t \in \Re^{k_2 \times k_2}$ is a diagonal matrix with $\mathbf{Q}^a{}_{(ii)} = 1$ if the ith class comes from the target data set, and $\mathbf{Q}^t{}_{(ii)} = 0$ otherwise. Note that, we need the both matrices because one class could appear in the both data sets. To encode the shared classes, we define one more indication matrix $\mathbf{Q} \in \Re^{k_2 \times k_2}$, which is a diagonal matrix with $\mathbf{Q}_{(ii)} = 1$ if the ith class is shared by the both languages, and $\mathbf{Q}_{(ii)} = 0$ otherwise:

$$\mathbf{Q}_{(ii)} = \begin{cases} 1 & \text{if } \mathbf{Q}^a{}_{(ii)} = 1 \text{ and } \mathbf{Q}^t{}_{(ii)} = 1, \\ 0 & \text{if } \mathbf{Q}^a{}_{(ii)} = 0 \text{ or } \mathbf{Q}^t{}_{(ii)} = 0. \end{cases} \tag{5}$$

We summarize the frequently used notations in Table I for convenience.

3.2. Objective of Our Proposed Approach

Given the Web page data $\mathbf{X}^a$ in the auxiliary language and their corresponding labels $\mathbf{Y}^a$, adopting the idea of simultaneous clustering of words and Web pages via NMTF, we may minimize the following objective [Gu and Zhou 2009b; Zhuang et al. 2010]:

$$\begin{aligned} J_{\mathrm{a}} &= \|\mathbf{X}^a - \mathbf{F}^a \mathbf{S}^a (\mathbf{G}^a)^T\|^2 + \alpha \, \mathbf{tr}[\mathbf{Q}^a (\mathbf{G}^a - \mathbf{Y}^a)^T \mathbf{C}^a (\mathbf{G}^a - \mathbf{Y}^a)], \\ s.t. \quad & \mathbf{F}^a \geq 0, \mathbf{S}^a \geq 0, \mathbf{G}^a \geq 0, (\mathbf{F}^a)^T \mathbf{F}^a = \mathbf{I}, (\mathbf{G}^a)^T \mathbf{G}^a = \mathbf{I}. \end{aligned} \tag{6}$$

Table I. Frequently used Notations in this Article

$\mathbf{X}^a$	data matrix of Web pages in auxiliary language
$\mathbf{X}^t$	data matrix of Web pages in target language
n^a	number of Web pages in auxiliary language
n^t	number of Web pages in target language
$\mathbf{F}^a$	word cluster indicator matrix of $\mathbf{X}^a$
$\mathbf{F}^t$	word cluster indicator matrix of $\mathbf{X}^t$
$\mathbf{S}^a$	the matrix associating word clusters and classes in the auxiliary language
$\mathbf{S}^t$	the matrix associating word clusters and classes in the target language
$\mathbf{G}^a$	class indicator matrix of Web pages in auxiliary language
$\mathbf{G}^t$	class indicator matrix of Web pages in target language
$\mathbf{Y}^a$	true label indicator matrix of Web pages in auxiliary language
$\mathbf{Y}^t$	true label indicator matrix of Web pages in target language
$\mathbf{V}$	word sharing indication matrix of the two data sets
$\mathbf{Q}$	class sharing indication matrix of the two data sets
$\mathbf{C}^a$	label indication matrix of $\mathbf{X}^a$
$\mathbf{C}^t$	label indication matrix of $\mathbf{X}^t$
k_1	number of word clusters
k_2	number of Web page classes

In Equation (6), $\alpha > 0$ is a parameter determining to which extent we enforce the prior labeling knowledge in auxiliary language, that is, $\mathbf{G}^a \approx \mathbf{Y}^a$. $\mathbf{Q}^a$ is used to control the scope to enforce the prior knowledge only to the classes that belong to the source data set.

Solving the optimization problem in Equation (6), we obtain the optimal $\mathbf{F}^{a*}$ and $\mathbf{S}^{a*}$, which contain the information of the Web page data in the auxiliary language. Our goal is therefore to transfer the learned knowledge encoded in $\mathbf{F}^{a*}$ and $\mathbf{S}^{a*}$ to the Web page data in the target language to improve the classification accuracy therein.

3.2.1. Transfer Knowledge via Word Clustering Approximation by $\mathbf{F}^a$ and $\mathbf{F}^t$. As illustrated earlier, we have two paths to transfer knowledge across the languages, among which the first one is achieved by minimizing the following optimization objective:

$$\begin{aligned} J_{\text{Trans-F}} = &\|\mathbf{X}^t - \mathbf{F}^t\mathbf{S}^t(\mathbf{G}^t)^T\|^2 + \beta\,\mathbf{tr}[\mathbf{Q}^t(\mathbf{G}^t - \mathbf{Y}^t)^T\mathbf{C}^t(\mathbf{G}^t - \mathbf{Y}^t)] \\ &+ \gamma\,\mathbf{tr}[(\mathbf{F}^t - \mathbf{F}^{a*})^T\mathbf{V}(\mathbf{F}^t - \mathbf{F}^{a*})], \\ &s.t. \quad \mathbf{F}^t \geq 0, \mathbf{S}^t \geq 0, \mathbf{G}^t \geq 0, (\mathbf{F}^t)^T\mathbf{F}^t = \mathbf{I}, (\mathbf{G}^t)^T\mathbf{G}^t = \mathbf{I}, \end{aligned} \tag{7}$$

where $\beta > 0$ and $\gamma > 0$ are two parameters. The second term of Equation (7) acts the same as that in Equation (6), which enforces labeling information in the target domain if it is available. The key part is the third term of Equation (7). It enforces the constraint that the word clusters in $\mathbf{X}^t$ are approximately close to $\mathbf{F}^a$, which is learned from $\mathbf{X}^a$. The extent of this approximation is determined by the parameter γ. As a result, the prior labeling information contained in $\mathbf{G}^a$ for $\mathbf{X}^a$ is transferred to the label assignments $\mathbf{G}^t$ for $\mathbf{X}^t$ via the semantic word structures $\mathbf{F}^a$ and $\mathbf{F}^t$, which is schematically shown as the red path in Figure 2.

In order to demonstrate the usefulness of KNWs in cross-language Web page classification, we give a more theoretical analysis on the example in Figure 1. Suppose that the knowledge in auxiliary language is certain, we may set γ in Equation (7) as ∞, which leads to the following objective to minimize:

$$J_2' = \|\mathbf{X}^t - \mathbf{F}^{a*}\mathbf{S}^t(\mathbf{G}^t)^T\|^2, \tag{8}$$

in which we temporarily ignore the training information in target language, in order to see the real effect of prior labeling knowledge to improve classification performance. The objective in Equation (8) is identical to the following optimization objective

[Ding et al. 2005, 2006]:

$$\max_{\mathbf{G}^t} \mathbf{tr}[(\mathbf{G}^t)^T(\mathbf{X}^t)^T\mathbf{F}^{a*}(\mathbf{F}^{a*})^T\mathbf{X}^t\mathbf{G}^t]. \tag{9}$$

By the equivalence between K-means clustering and principal component analysis (PCA) [Ding and He 2004; Zha et al. 2001], the clustering by Equation (9) uses $(\mathbf{X}^t)^T\mathbf{F}^{a*}(\mathbf{F}^{a*})^T\mathbf{X}^t$ as the pairwise similarity, whereas K-means clustering uses $(\mathbf{X}^t)^T\mathbf{X}^t$ as the pairwise similarity. For example in Figure 1, we have

$$(\mathbf{X}^t)^T\mathbf{X}^t = \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix}. \tag{10}$$

Thus K-means clustering will produce (W1, W3) as a cluster and (W2, W4) as another cluster.

Now, with the work pattern knowledge $\mathbf{F}^{a*}$ learned from the auxiliary language, we have

$$(\mathbf{X}^t)^T\mathbf{F}^{a*}(\mathbf{F}^{a*})^T\mathbf{X}^t = \begin{bmatrix} 1 & 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 \end{bmatrix}, \tag{11}$$

in which we assume we already learned $\mathbf{F}^{a*}$ from auxiliary language, which, as shown in Figure 1(c), is

$$(\mathbf{F}^{a*})^T = 2^{-1/2}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{12}$$

Clearly, using the similarity in Equation (11), K-means clustering will generate (W1, W2) as a cluster and (W3, W4) as another cluster, which is a more meaningful result as in Figure 1(c).

3.2.2. Transfer Knowledge via the Associations Between the Word Clusters and Web Page Classes by $\mathbf{S}^a$ and $\mathbf{S}^t$. As discussed earlier in Section 1, compared to word clusters, the associations between word clusters and Web pages classes are more reliable to convey semantic relationships across different languages. Formally, we achieve this by minimizing the following optimization objective:

$$\begin{aligned} J_{\text{Trans-S}} &= \|\mathbf{X}^t - \mathbf{F}^t\mathbf{S}^t(\mathbf{G}^t)^T\|^2 + \mu\,\mathbf{tr}[(\mathbf{S}^{a*} - \mathbf{S}^t)\mathbf{Q}(\mathbf{S}^{a*} - \mathbf{S}^t)^T] \\ &s.t. \quad \mathbf{F}^t \geq 0, \mathbf{S}^t \geq 0, \mathbf{G}^t \geq 0, (\mathbf{F}^t)^T\mathbf{F}^t = \mathbf{I}, (\mathbf{G}^t)^T\mathbf{G}^t = \mathbf{I}, \end{aligned} \tag{13}$$

where $\mathbf{S}^{a*}$ is obtained by solving Equation (6). As a result, $\mathbf{S}^{a*}$, learned from the auxiliary data set, is used as the supervision to classify the target data. Namely, $\mathbf{S}^{a*}$ and $\mathbf{S}^t$ bridge the source and target languages such that prior labeling knowledge can be transferred from the former to the latter, which is schematically shown as the green path in Figure 2.

A simpler way to transfer the supervised knowledge via the associations between word patterns and semantic classes is to force the middle factor matrix $\mathbf{S}$ in the two data matrix factorizations for the both domains to be identical, which is achieved by minimize the following objective:

$$\begin{aligned} J'_{\text{Trans-S}} &= \|\mathbf{X}^t - \mathbf{F}^t\mathbf{S}^{a*}(\mathbf{G}^t)^T\|^2 \\ &s.t. \quad \mathbf{F}^t \geq 0, \mathbf{G}^t \geq 0, (\mathbf{F}^t)^T\mathbf{F}^t = \mathbf{I}, (\mathbf{G}^t)^T\mathbf{G}^t = \mathbf{I}. \end{aligned} \tag{14}$$

Because Equation (14) is free of parameter, it easier to fine tune in practice, which thereby is used in our earlier conference publication [Wang et al. 2011b, 2011e] to transfer the supervised knowledge across languages. However, compared to Equation (13), Equation (14) is less general and flexible for the following two reasons.

First, instead of rigidly forcing the two matrix factorizations in the both domains to share an identical middle factor matrix, through the parameter μ Equation (13) allows user to adjust to which extent we approximate $\mathbf{S}^a$ by $\mathbf{S}^t$. If it is known in advance that the auxiliary data has very close distribution to the target data, we may set μ to be a large value, such that the factorization in the target domain is strongly guided by the supervised knowledge in the auxiliary domain. Apparently, when μ is sufficiently large, Equation (13) is exactly equivalent to Equation (14). In contrast, if the auxiliary data is known in *a priori* to have a very different distribution from that of the target data, we may set μ to be a small value so as not to bias the factorization in the target domain too much, while the useful knowledge in the auxiliary domain is still used.

Second, but more important, Equation (13) could potentially reduce the noise due to semantic class differences. Because Equation (13) only approximate $\mathbf{S}^t$ to $\mathbf{S}^a$ for the classes shared by the both domains due to the introduction of class sharing indication matrix $\mathbf{Q}$, when the auxiliary data and target data do not share a same set of classes, the supervised information for the classes exclusively in the auxiliary data will have no effect on the data factorization in the target domain. On the other hand, when the data in the both domains convey same semantic meanings by sharing a same set of classes, one can set $\mathbf{Q} = \mathbf{I}$, which reduces Equation (13) to Equation (14) when μ is sufficiently large.

In summary, both Equation (13) and Equation (14) are able to transfer the prior labeling information in the auxiliary data to the target data, while the former is more general and flexible. Equation (14) can be seen as a special case of Equation (13). Therefore, in the current work, we choose Equation (13) as our objective to build the path to transfer supervised knowledge.

3.2.3. Our Optimization Objective. Finally, we may combine the three optimization problems in Equation (6), Equation (7) and Equation (13) into a joint optimization objective to minimize the following objective:

$$\begin{aligned} J_{\mathrm{DKT}} = & \|\mathbf{X}^a - \mathbf{F}^a\mathbf{S}^a(\mathbf{G}^a)^T\|^2 + \|\mathbf{X}^t - \mathbf{F}^t\mathbf{S}^t(\mathbf{G}^t)^T\|^2 \\ & + \alpha\,\mathbf{tr}\left[\mathbf{Q}^a(\mathbf{G}^a - \mathbf{Y}^a)^T\mathbf{C}^a(\mathbf{G}^a - \mathbf{Y}^a)\right] \\ & + \beta\,\mathbf{tr}\left[\mathbf{Q}^t(\mathbf{G}^t - \mathbf{Y}^t)^T\mathbf{C}^t(\mathbf{G}^t - \mathbf{Y}^t)\right] \\ & + \gamma\,\mathbf{tr}\left[(\mathbf{F}^t - \mathbf{F}^a)^T\mathbf{V}(\mathbf{F}^t - \mathbf{F}^a)\right], \\ & + \mu\,\mathbf{tr}\left[(\mathbf{S}^a - \mathbf{S}^t)\mathbf{Q}(\mathbf{S}^a - \mathbf{S}^t)^T\right] \\ s.t. \quad & \mathbf{F}^a \geq 0, \mathbf{G}^a \geq 0, \mathbf{S}^a \geq 0, \mathbf{S}^t \geq 0, \mathbf{F}^t \geq 0, \mathbf{G}^t \geq 0, \\ & (\mathbf{F}^a)^T\mathbf{F}^a = \mathbf{I}, (\mathbf{G}^a)^T\mathbf{G}^a = \mathbf{I}, (\mathbf{F}^t)^T\mathbf{F}^t = \mathbf{I}, (\mathbf{G}^t)^T\mathbf{G}^t = \mathbf{I}. \end{aligned} \tag{15}$$

In this formulation, we approximate $\mathbf{S}^t$ and $\mathbf{F}^t$ in the target domain to those in the source domain, which are used as the two bridges to transfer knowledge between them as illustrated in Figure 2.

Note that, the last term of Equation (7) only applies to the common words of $\mathbf{X}^a$ and $\mathbf{X}^t$, which are encoded by $\mathbf{V}$. When the auxiliary data set and the target data set do not share any word, that is, $\mathbf{V} = 0^{m\times m}$ is a zero matrix, there will be no knowledge transfer through word path. Similarly, if the auxiliary data set and the target data set do not share common classes, there will be no knowledge transformation in the optimization problem of Equation (13), because it is decoupled into two independent subproblems,

one for auxiliary data and the other for target data. However, these two cases rarely happen at the same time. As a result, our model is flexible and can always transfer knowledge in Equation (15) through either word clusters or the associations between word clusters and Web page classes, or the both.

On solving Equation (15), there exist a number of ways to determine the class labels of unlabeled Web pages in target language. In this work, following [Ding et al. 2005], we consider $\mathbf{g}_{i\cdot}^t$ (after normalization) as the posterior probability of class membership, we assign the class label to the $\mathbf{x}_{\cdot i}^t$ in target language using the following rule:

$$l\left(\mathbf{x}_{\cdot i}^t\right) = \arg\max_k \ \mathbf{G}_{(ik)}^t. \tag{16}$$

Solving Equation (15) and assigning labels to the unlabeled Web pages in target language using Equation (16), we can classify cross-language Web pages in the target language domain. Because Equation (15) transfers knowledge in two different paths, we call it as the proposed *DKT* approach.

3.3. Optimization Algorithm

In the rest of this section, we will derive the solution to Equation (15) and present an alternating scheme to optimize the objective J_{DKT}. Specifically, we will optimize one variable while fixing the rest variables. The procedure repeats until convergence.

First, we expand the objective in Equation (15) as follows,

$$\begin{aligned}
&J(\mathbf{F}^a, \mathbf{S}^a, \mathbf{G}^a, \mathbf{F}^t, \mathbf{S}^t, \mathbf{G}^t)\\
&\quad = \mathbf{tr}[-2(\mathbf{X}^a)^T\mathbf{F}^a\mathbf{S}^a(\mathbf{G}^a)^T + \mathbf{G}^a(\mathbf{S}^a)^T(\mathbf{F}^a)^T\mathbf{F}^a\mathbf{S}^a(\mathbf{G}^a)^T]\\
&\quad \mathbf{tr}[-2(\mathbf{X}^t)^T\mathbf{F}^t\mathbf{S}^t(\mathbf{G}^t)^T + \mathbf{G}^t(\mathbf{S}^t)^T(\mathbf{F}^t)^T\mathbf{F}^t\mathbf{S}^t(\mathbf{G}^t)^T]\\
&\quad + \alpha\,\mathbf{tr}[\mathbf{Q}^a(\mathbf{G}^a)^T\mathbf{C}^a\mathbf{G}^a - 2\mathbf{Q}^a(\mathbf{G}^a)^T\mathbf{C}^a\mathbf{Y}^a]\\
&\quad + \beta\,\mathbf{tr}[\mathbf{Q}^t(\mathbf{G}^t)^T\mathbf{C}^t\mathbf{G}^t - 2\mathbf{Q}^t(\mathbf{G}^t)^T\mathbf{C}^t\mathbf{Y}^t]\\
&\quad + \gamma\,\mathbf{tr}[(\mathbf{F}^t)^T\mathbf{V}\mathbf{F}^t - 2(\mathbf{F}^t)^T\mathbf{V}\mathbf{F}^a + (\mathbf{F}^a)^T\mathbf{V}\mathbf{F}^a]\\
&\quad + \mu\,\mathbf{tr}[\mathbf{Q}(\mathbf{S}^a)^T\mathbf{S}^a - 2\mathbf{Q}(\mathbf{S}^a)^T\mathbf{S}^t + \mathbf{Q}(\mathbf{S}^t)^T\mathbf{S}^t],
\end{aligned} \tag{17}$$

in which constant terms are discarded.

3.3.1. Computation of $\mathbf{F}^a$. We first compute $\mathbf{F}^a$ and assume the rest variables are fixed. Following the standard theory of constrained optimization, we introduce the Lagrangian multipliers $\mathbf{U} \in \Re^{k_1 \times k_1}$ (a symmetric matrix of size $k_1 \times k_1$) and minimize the following Lagrangian function:

$$L(\mathbf{F}^a) = J - \mathbf{tr}[\mathbf{U}((\mathbf{F}^a)^T\mathbf{F}^a - \mathbf{I})]. \tag{18}$$

Thus, the gradient of L is:

$$\frac{\partial L}{\partial \mathbf{F}^a} = -2\mathbf{X}^a\mathbf{G}^a(\mathbf{S}^a)^T + 2\mathbf{F}^a\mathbf{S}^a(\mathbf{G}^a)^T\mathbf{G}^a(\mathbf{S}^a)^T - 2\gamma\mathbf{V}\mathbf{F}^t + 2\gamma\mathbf{V}\mathbf{F}^a + 2\mathbf{F}^a\mathbf{U}. \tag{19}$$

The Karush–Kuhn–Tucker (KKT) condition complementarity condition gives

$$(-2\mathbf{X}^a\mathbf{G}^a(\mathbf{S}^a)^T + 2\mathbf{F}^a\mathbf{S}^a(\mathbf{G}^a)^T\mathbf{G}^a(\mathbf{S}^a)^T - 2\gamma\mathbf{V}\mathbf{F}^t + 2\gamma\mathbf{V}\mathbf{F}^a + 2\mathbf{F}^a\mathbf{U})_{ik}(\mathbf{F}^a)_{ik} = 0. \tag{20}$$

This is the fixed point relation that local minima for $\mathbf{F}^a$ must hold.

The standard approach is to solve the coupled equations Equation (20) and constraint $\mathbf{F}^a\mathbf{F}^a = \mathbf{I}$ for $\mathbf{F}^a$ and $\mathbf{U}$. This system nonlinear equations is generally difficult to solve. Following [Ding et al. 2006, Section 6], we can derive the Lagrangian multiplier $\mathbf{U}$ is

computed as following:

$$\mathbf{U} = (\mathbf{F}^a)^T \mathbf{X}^a \mathbf{G}^a (\mathbf{S}^a)^T - \mathbf{S}^a (\mathbf{G}^a)^T \mathbf{G}^a (\mathbf{S}^a)^T . \tag{21}$$

Following the same derivations in Ding et al. [2005, 2006, 2010], we can obtain updating formula as follows:

$$\mathbf{F}^a_{(ij)} \leftarrow \mathbf{F}^a_{(ij)} \sqrt{\frac{(\mathbf{X}^a \mathbf{G}^a (\mathbf{S}^a)^T + \gamma \mathbf{V}\mathbf{F}^t)_{(ij)}}{(\mathbf{F}^a (\mathbf{F}^a)^T \mathbf{X}^a \mathbf{G}^a (\mathbf{S}^a)^T + \gamma \mathbf{V}\mathbf{F}^a)_{(ij)}}}. \tag{22}$$

Based on the updating formulation in Equation (22) earlier, it is obvious that the constraint $\mathbf{F}^a \geq 0$ is automatically satisfied.

3.3.2. Computation of $\mathbf{S}^a$, $\mathbf{G}^a$, $\mathbf{F}^t$, $\mathbf{S}^t$ *and* $\mathbf{G}^t$. Following the same derivations in Equations (18–22), we obtain the updating rules for the rest variables of J_{DKT} as following:

$$\mathbf{G}^a_{(ij)} \leftarrow \mathbf{G}^a_{(ij)} \sqrt{\frac{((\mathbf{X}^a)^T \mathbf{F}^a \mathbf{S}^a + \alpha \mathbf{C}^a \mathbf{Y}^a \mathbf{Q}^a)_{(ij)}}{(\mathbf{G}^a (\mathbf{G}^a)^T (\mathbf{X}^a)^T \mathbf{F}^a \mathbf{S}^a + \alpha \mathbf{C}^a \mathbf{G}^a \mathbf{Q}^a)_{(ij)}}} \tag{23}$$

$$\mathbf{S}^a_{ij} \leftarrow \mathbf{S}^a_{ij} \sqrt{\frac{((\mathbf{F}^a)^T \mathbf{X}^a \mathbf{G}^a + \mu \mathbf{Q}\mathbf{S}^t)_{(ij)}}{((\mathbf{F}^a)^T \mathbf{F}^a \mathbf{S}^a (\mathbf{G}^a)^T \mathbf{G}^a + \mu \mathbf{Q}\mathbf{S}^a)_{(ij)}}} \tag{24}$$

$$\mathbf{F}^t_{(ij)} \leftarrow \mathbf{F}^t_{(ij)} \sqrt{\frac{(\mathbf{X}^t \mathbf{G}^t (\mathbf{S}^t)^T + \gamma \mathbf{V}\mathbf{F}^a)_{(ij)}}{(\mathbf{F}^t (\mathbf{F}^t)^T \mathbf{X}^t \mathbf{G}^t (\mathbf{S}^t)^T + \gamma \mathbf{V}\mathbf{F}^t)_{(ij)}}} \tag{25}$$

$$\mathbf{G}^t_{(ij)} \leftarrow \mathbf{G}^t_{(ij)} \sqrt{\frac{((\mathbf{X}^t)^T \mathbf{F}^t \mathbf{S}^t + \beta \mathbf{C}^t \mathbf{Y}^t \mathbf{Q}^t)_{(ij)}}{(\mathbf{G}^t (\mathbf{G}^t)^T (\mathbf{X}^t)^T \mathbf{F}^t \mathbf{S}^t + \beta \mathbf{C}^t \mathbf{G}^t \mathbf{Q}^t)_{(ij)}}} \tag{26}$$

$$\mathbf{S}^t_{ij} \leftarrow \mathbf{S}^t_{ij} \sqrt{\frac{((\mathbf{F}^t)^T \mathbf{X}^t \mathbf{G}^t + \mu \mathbf{Q}\mathbf{S}^a)_{(ij)}}{((\mathbf{F}^t)^T \mathbf{F}^t \mathbf{S}^t (\mathbf{G}^a)^T \mathbf{G}^t + \mu \mathbf{Q}\mathbf{S}^t)_{(ij)}}} \tag{27}$$

In summary, we present an iterative multiplicative updating algorithm to optimize Equation (15) in Algorithm 1.

The analysis of the convergence of Algorithm 1 is provided in the appendix.

4. FAST DUAL KNOWLEDGE TRANSFER FOR LARGE-SCALE WEB DATA

Despite its mathematical elegance, same as existing NMF based methods [Ding et al. 2005, 2006, 2010; Gu and Zhou 2009a; Li et al. 2010; Wang et al. 2008], Equation (15) suffers from two problems that impede its practical use. First, similar to Equation (2), relaxing $\mathbf{F}$ and $\mathbf{G}$ to be continuous variables makes the immediate outputs of Equation (15) not the clustering labels, which requires an additional post-processing step, for example, using Equation (16), and may lead to non-unique solutions. Second, and more important, Equation (15) is solved by an alternately iterative algorithm as described in Algorithm 1, and in each iteration step the intensive matrix multiplications are involved. As a result, it is infeasible to apply such algorithms to large-scale real world Web data due to the expensive computational cost.

ALGORITHM 1: Algorithm to solve J_{DKT} in Equation (15).

Input: 1. Data matrix $\mathbf{X}^a$ in auxiliary language,
2. data matrix $\mathbf{X}^t$ in target language,
3. labels of Web pages in auxiliary language $\mathbf{Y}^a$,
4. optional labeling information $\mathbf{Y}^t$ in target data,
5. trade-off parameters α, β and γ.
Initialize $\mathbf{F}^a$, $\mathbf{G}^a$, $\mathbf{S}^a$, $\mathbf{F}^t$, $\mathbf{G}^t$ and $\mathbf{S}^t$ following [Zhuang et al. 2010];
while *not converge* **do**
1. Update $\mathbf{F}^a$ using Equation (22),
2. Update $\mathbf{G}^a$ using Equation (23),
3. Update $\mathbf{S}^a$ using Equation (24),
4. Update $\mathbf{F}^t$ using Equation (25),
5. Update $\mathbf{G}^t$ using Equation (26),
6. Update $\mathbf{S}^t$ using Equation (27),

end
Predict labels for $\mathbf{x}^t_{\cdot i}$ using Equation (16).
Output: Labels assigned to the labeled Web page $\mathbf{x}^t_{\cdot i}$ in target language.

In order to tackle these difficulties, instead of solving the relaxed clustering problems as in Equations (2)–(4) and Equation (15), we solve the original clustering problem similar to Equation (1). Specifically, we constrain the factor matrices of NMTF to be cluster indicator matrices and minimize the following objective:

$$\begin{aligned} J_{\text{F-DKT}} &= \|\mathbf{X}^a - \mathbf{F}^a\mathbf{S}^a(\mathbf{G}^a)^T\|^2 + \|\mathbf{X}^t - \mathbf{F}^t\mathbf{S}^t(\mathbf{G}^t)^T\|^2 \\ &\quad + \alpha\,\mathbf{tr}\left[\mathbf{Q}^a(\mathbf{G}^a - \mathbf{Y}^a)^T\mathbf{C}^a(\mathbf{G}^a - \mathbf{Y}^a)\right] \\ &\quad + \beta\,\mathbf{tr}\left[\mathbf{Q}^t(\mathbf{G}^t - \mathbf{Y}^t)^T\mathbf{C}^t(\mathbf{G}^t - \mathbf{Y}^t)\right] \\ &\quad + \gamma\,\mathbf{tr}\left[(\mathbf{F}^t - \mathbf{F}^a)^T\mathbf{V}(\mathbf{F}^t - \mathbf{F}^a)\right], \\ &\quad + \mu\,\mathbf{tr}\left[(\mathbf{S}^a - \mathbf{S}^t)\mathbf{Q}(\mathbf{S}^a - \mathbf{S}^t)^T\right] \\ &\quad s.t. \quad \mathbf{F}^a \in \Psi,\ \mathbf{G}^a \in \Psi,\ \mathbf{S}^a \geq 0,\ \mathbf{S}^t \geq 0,\ \mathbf{F}^t \in \Psi,\ \mathbf{G}^t \in \Psi. \end{aligned} \tag{28}$$

We call Equation (28) as the proposed F-DKT. Note that, in Equation (28) the traditional nonnegative constraints on the four factor matrices in the two matrix factorizations are replaced by cluster indicator matrices. Because $\Psi^{n\times c} \subset \Re_+^{n\times c}$, these new constraints are more stringent. Surprisingly, with these new constraints, though more stringent, as shown theoretically shortly in this section and empirically later in Section 6, the computational speed of our approach can be significantly improved.

4.1. Optimization Procedures

Again, we alternately optimize the five variables of $J_{\text{F-DKT}}$ in Equation (28).

First, when $\mathbf{F}^a$, $\mathbf{G}^a$, $\mathbf{F}^t$, and $\mathbf{G}^t$ are fixed, the optimizations of $\mathbf{S}^a$ and $\mathbf{S}^t$ are same as before, therefore the same updating rule in Equation (24) and Equation (27) are used.

Second, when $\mathbf{F}^a$, $\mathbf{S}^a$, $\mathbf{F}^t$, $\mathbf{S}^t$, and $\mathbf{G}^t$ are fixed, the optimization problem to obtain $\mathbf{G}^a$ can be decoupled and we solve the following simpler problem for each i $(1 \leq i \leq n^a)$:

$$\min_{\mathbf{G}^a\in\Psi} \left\|\mathbf{x}^a_{\cdot i} - \mathbf{F}^a\mathbf{S}^a\left(\mathbf{g}^a_{i\cdot}\right)^T\right\|^2 + \alpha\,\mathbf{C}^a_{(ii)}\left(\mathbf{g}^a_{i\cdot} - \mathbf{y}^a_{i\cdot}\right)\mathbf{Q}^a\left(\mathbf{g}^a_{i\cdot} - \mathbf{y}^a_{i\cdot}\right)^T. \tag{29}$$

Because $\mathbf{g}^a_{i\cdot}\,(1 \leq i \leq n^a) \in \Psi^{1\times k_2}$ is a cluster indicator vector in which one and only one element is 1 and the rest are zeros, the solution of Equation (29) can be easily

obtained by:

$$\mathbf{G}^a_{(ij)} = \begin{cases} 1 & j = \arg\min_k \ \left\|\mathbf{x}^a_{\cdot i} - \tilde{\mathbf{f}}^a_{\cdot k}\right\|^2 + \alpha \mathbf{C}^a_{(ii)} \mathbf{Q}^a{}_{(kk)} \left(1 - \mathbf{Y}^a_{(ik)}\right)^2, \\ 0 & \text{otherwise,} \end{cases} \tag{30}$$

where $\tilde{\mathbf{F}}^a = \mathbf{F}^a \mathbf{S}^a$ and $\tilde{\mathbf{f}}^a_{\cdot k}$ is the kth column of $\mathbf{F}^a$. Note that Equation (30) simply enumerates the k_2 vector norms and seeks the minimum one, without involving any matrix multiplication.

Similarly, when $\mathbf{G}^a$, $\mathbf{S}^a$, $\mathbf{F}^t$, $\mathbf{S}^t$, and $\mathbf{G}^t$ are fixed, the optimization problem to obtain $\mathbf{F}^a$ can be decoupled and we solve the following simpler problem for each $j\,(1 \le j \le d)$:

$$\min_{\mathbf{F}^a \in \Psi} \left\|\mathbf{x}^a_{j\cdot} - \mathbf{f}^a_{j\cdot} \mathbf{S}^a (\mathbf{G}^a)^T\right\|^2 + \gamma \, \mathbf{V}_{(ii)} \left\|\mathbf{f}^a_{j\cdot} - \mathbf{f}^t_{j\cdot}\right\|^2. \tag{31}$$

Again, because $\mathbf{f}^a_{j\cdot}\,(1 \le j \le d) \in \Psi^{1 \times k_1}$ is a cluster indicator vector for feature side, the solution to Equation (31) is

$$\mathbf{F}^a_{(ij)} = \begin{cases} 1 & i = \arg\min_l \ \left\|\mathbf{x}^a_{j\cdot} - \tilde{\mathbf{g}}^a_{l\cdot}\right\|^2 + \gamma \mathbf{V}_{(ii)} \left(1 - \mathbf{F}^t_{(jl)}\right)^2, \\ 0 & \text{otherwise,} \end{cases} \tag{32}$$

where $(\tilde{\mathbf{G}}^a)^T = \mathbf{S}^a (\mathbf{G}^a)^T$ and $\tilde{\mathbf{g}}^a_{l\cdot}$ is the lth row of $(\tilde{\mathbf{G}}^a)^T$.

Following the same idea, we can obtain the update rules for $\mathbf{F}^t$ and $\mathbf{G}^t$. Specifically, when $\mathbf{F}^a$, $\mathbf{G}^a$, $\mathbf{S}^a$, $\mathbf{F}^t$, and $\mathbf{S}^t$ are fixed, the optimization problem to obtain $\mathbf{G}^t$ can be decoupled and we solve the following simpler problem for each $i (1 \le i \le n^t)$:

$$\min_{\mathbf{G}^t \in \Psi} \ \left\|\mathbf{x}^t_{\cdot i} - \mathbf{F}^t \mathbf{S}^t \left(\mathbf{g}^t_{i\cdot}\right)^T\right\|^2 + \alpha \ \mathbf{C}^t_{(ii)} \left(\mathbf{g}^t_{i\cdot} - \mathbf{y}^t_{i\cdot}\right) \mathbf{Q}^t \left(\mathbf{g}^t_{i\cdot} - \mathbf{y}^t_{i\cdot}\right)^T. \tag{33}$$

Thus, let $\tilde{\mathbf{F}}^t = \mathbf{F}^t \mathbf{S}^t$, the solution of Equation (33) can be obtained by

$$\mathbf{G}^t_{(ij)} = \begin{cases} 1 & j = \arg\min_k \ \left\|\mathbf{x}^t_{\cdot i} - \tilde{\mathbf{f}}^t_{\cdot k}\right\|^2 + \alpha \mathbf{C}^t_{(ii)} \mathbf{Q}^t{}_{(kk)} \left(1 - \mathbf{Y}^t_{(ik)}\right)^2, \\ 0 & \text{otherwise.} \end{cases} \tag{34}$$

Finally, when $\mathbf{F}^a$, $\mathbf{G}^a$, $\mathbf{S}^a$, $\mathbf{S}^t$ and $\mathbf{G}^t$ are fixed, the optimization problem to obtain $\mathbf{F}^t$ can be decoupled and we solve the following simpler problem for each $j\,(1 \le j \le d)$:

$$\min_{\mathbf{F}^t \in \Psi} \left\|\mathbf{x}^t_{j\cdot} - \mathbf{f}^t_{j\cdot} \mathbf{S}^t (\mathbf{G}^t)^T\right\|^2 + \gamma \, \mathbf{V}_{(ii)} \left\|\mathbf{f}^t_{j\cdot} - \mathbf{f}^a_{j\cdot}\right\|^2. \tag{35}$$

Let $(\tilde{\mathbf{G}}^t)^T = \mathbf{S}^t (\mathbf{G}^t)^T$, the solution to Equation (35) is

$$\mathbf{F}^t_{(ij)} = \begin{cases} 1 & i = \arg\min_l \ \left\|\mathbf{x}^t_{j\cdot} - \tilde{\mathbf{g}}^t_{l\cdot}\right\|^2 + \gamma \mathbf{V}_{(ii)} \left(1 - \mathbf{F}^a_{(jl)}\right)^2, \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

The procedures to optimize $J_{\text{F-DKT}}$ in Equation (28) are summarized in Algorithm 2. Due to the nature of alternating optimization, Algorithm 2 is guaranteed to converge to a local minima (existing NMF algorithms [Ding et al. 2005, 2006, 2010] also converges to a local minima because the objectives J_{NMF} and J_{NMTF} are not convex in both variables $\mathbf{F}$ and $\mathbf{G}$).

A careful look at Algorithm 2 shows that the steps 3–6 are obtained by enumerating vector norms, which is definitely much more computationally efficient than matrix multiplications used in the same steps of Algorithm 1. Because the updates on $\mathbf{S}^a$ and $\mathbf{S}^t$ is performed on much smaller matrices (typically $k_1 \ll d$, $k_2 \ll n^a$ and $k_2 \ll n^t$), the main computational loads of the algorithms fall on steps 3–6. As a result, the computational speed of Algorithm 2 is much faster than Algorithm 1.

Moveover, on solution, $\mathbf{G}$ directly gives the classification results of data points, while additional postprocessing step is required in continuous NMTF based methods, such as using Equation (16) to extract labeling structures from the results of Equation (15).

ALGORITHM 2: Algorithm to solve $J_{\text{F-DKT}}$ in Equation (28).

Input: 1. Data matrix $\mathbf{X}^a$ in auxiliary language,
2. data matrix $\mathbf{X}^t$ in target language,
3. labels of Web pages in auxiliary language $\mathbf{Y}^a$,
4. optional labeling information $\mathbf{Y}^t$ in target data,
5. trade-off parameters α, β and γ.
Initialize $\mathbf{F}^a$, $\mathbf{S}^a$, $\mathbf{G}^a$, $\mathbf{F}^t$, smt and $\mathbf{G}^t$ following [Zhuang et al. 2010];
while *not converge* **do**
1. Update $\mathbf{S}^a$ using Equation (24),
2. Update $\mathbf{S}^t$ using Equation (27),
3. Update $\mathbf{G}^a$ using Equation (30),
4. Update $\mathbf{F}^a$ using Equation (32),
5. Update $\mathbf{G}^t$ using Equation (34),
6. Update $\mathbf{F}^t$ using Equation (36),

end
Output: Indicator matrix $\mathbf{G}^t$ for the class memberships of the unlabeled Web pages in target language.

5. RELATED WORKS

In this section, we review several prior researches mostly related to our work, including transfer learning, cross-language classification and NMTF.

5.1. Transfer Learning

From machine learning perspective of view, our work belongs to the topic of *transfer learning* (also called as *domain adaption* in some research papers), which deals with the case in which the training and test data are obtained from different resources thereby in different distributions [Ling et al. 2008; Li et al. 2009, 2010; Olsson et al. 2005; Prettenhofer and Stein 2010; Ramírez-de-la Rosa et al. 2010; Shi et al. 2010; Wan 2009; Wu and Lu 2008; Zhuang et al. 2010]. Recently, Dai et al. [2007] studied transfer learning through co-clustering, which also uses coclustering to achieve knowledge transfer across domain, same as ours. For a comprehensive survey of transfer learning, we refer readers to Pan and Yang [2009].

5.2. Cross-Language Classification

Cross-language Web page and document classification has attracted increased attention in recent years due to its importance in IR. Bel et al. [2003] studied English–Spanish cross-language classification problem. Two scenarios are considered in their work. One scenario assumes to have training documents in both languages, and the other is to learn a model from the text in one language and classify the data in another language by translation. Our work follows the second strategy. Olsson et al. [2005] employed a general probabilistic English-Czech dictionary to translate Czech text into English and then classified Czech documents using the classifier built on English training data. Ling et al. [2008] classify Chinese Web pages using English data source by utilizing the information bottleneck (IB) theory. Other cross-language text classification researches include [Wu and Lu 2008] (Chinese–English), Ramírez-de-la Rosa et al. [2010] (English–Spanish–French), Shi et al. [2010] (English–Chinese–French), and so forth, to be mentioned.

5.3. NMTF

NMF is a useful learning method to approximate a nonnegative input data matrix by the product of factor matrices [Lee and Seung 1999, 2001], which has been applied to solve many real world problems including dimensionality reduction, pattern

recognition, clustering and classification [Ding et al. 2010, 2005, 2006; Wang et al. 2008; Gu and Zhou 2009a; Chen et al. 2009; Li et al. 2009, 2010; Zhuang et al. 2010]. Recently, Ding *et al.* extended NMF [Ding et al. 2006] to NMTF and explored its relationships to K-means/spectral clustering [Ding et al. 2005, 2006]. Due to its mathematical elegance and encouraging empirical results, NMTF method is further developed to address a variety of aspects of unsupervised and semi-supervised learning [Chen et al. 2009; Gu and Zhou 2009a; Li et al. 2009, 2010; Wang et al. 2008; Zhuang et al. 2010], among which [Li et al. 2009] and [Zhuang et al. 2010] are closely related to our work. The former investigated cross-domain sentiment classification, which transfers knowledge by sharing information of word clusters. This is similar to our approach to transfer knowledge through words. While they dealt with two separate tasks of matrix factorizations, first on the source domain and then on the target domain, our approach optimizes a combined and collaborative objective, which leads to extra values in classification as shown later in our experimental evaluations. In addition, they assume there exist no label information in target domain, which restricts its capability to solve real world problems. The latter considered the cross-domain document classification via transferring knowledge by the associations between word clusters and document classes, which, however, did not use the important information contained in words as both our approach and [Li et al. 2009]. Again, they restrict that the data in the source domain are completely labeled while no data labeling information in the target domain. In summary, our approach has very close relationships to Li et al. [2009] and Zhuang et al. [2010], but enjoys the advantages of both of them, with additional flexibilities to allow training data appearing in various forms.

Most importantly, all these earlier NMTF based clustering or classification methods rely on solution algorithms involving intensive matrix multiplications, which makes them computational inefficient and can scale to large-scale real world data. In contrast, the fast implementation of our approach address this problem, which adds to its practical value.

6. EXPERIMENTS

In this section, we evaluate the proposed joint NMTF based DKT approach as well as its fast variant in cross-language Web page classification tasks.

6.1. Data Preparation

We conduct our empirical evaluations on a publicly available multi-lingual Web page data set—cross lingual sentiment corpus[5] [Prettenhofer and Stein 2010]. This data set contains about 800,000 web pages from Amazon web site for product reviews in four languages: English, German, French and Japanese. The crawled part of the corpus contains more than four millions of Web pages in the three languages other than English from amazon.{de|fr|co.jp}. Besides the original Web pages, all the Web pages in German, French and Japanese are translated into English. The corpus is extended with English Web pages provided by Blitzer et al. [2006]. All the Web pages in the corpus are divided into three categories on the product they describe: books, DVDs and music. We refer readers to Prettenhofer and Stein [2010] for the details of the data set and language translation procedures.

We randomly pick up 5,000 Web pages from each language in our test. Same as [Prettenhofer and Stein 2010], we use English as the auxiliary language and the rest three as target languages separately. Therefore we end up with three language pairs

[5]http://www.webis.de/research/corpora/.

Table II. Shared Keywords of the Three Language Pairs

Language pair	# shared keywords	Several sample shared keywords	
		English	German/French/Japanese
English–Germen	5012	absurd	absurd
		blew	blies
		darling	Liebling
		worst	schlimmsten
English–French	5608	abstract	résumé
		beginning	début
		decade	décennie
		work	travailler
English–Japanese	5605	5th	第5回
		aim	目標
		chose	選択
		familiarity	知識

Table III. Description of Testing Data Sets. English is used as Auxiliary Language in all the Testing Data Sets

Data	Target language	# Labeled Auxiliary	# Labeled Target
$D1$	German	3,500	0
$D2$	German	3,500	1,000
$D3$	French	3,500	0
$D4$	French	3,500	1,000
$D5$	Japanese	3,500	0
$D6$	Japanese	3,500	1,000

for testing: English-Germen, English-French, and English-Japanese. The numbers of shared keywords between the three language pairs are reported in Table II, in which several sample shared keywords are also shown for reference.

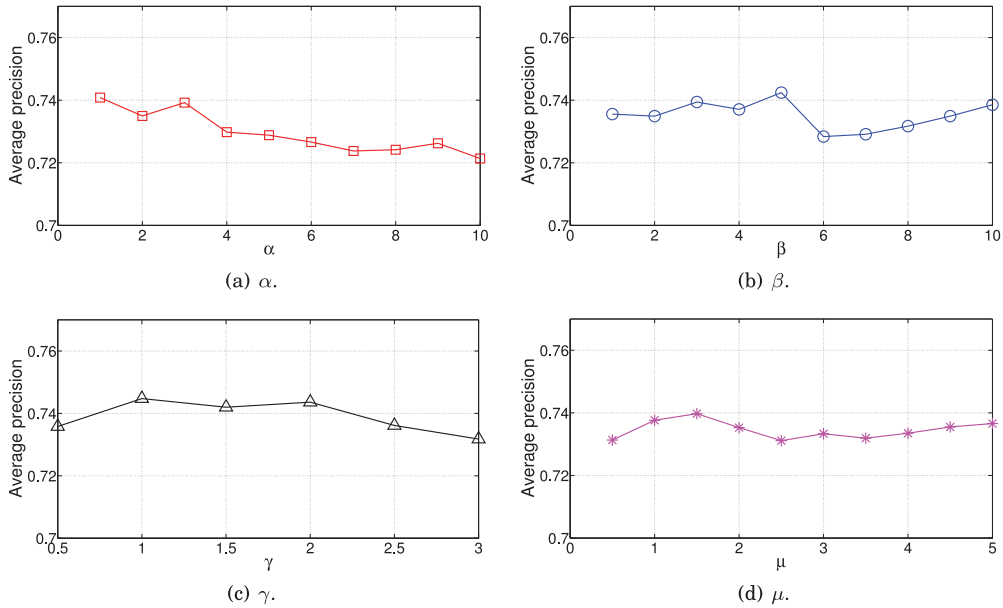
Because in real world applications not all the Web pages in the auxiliary language are labeled, we randomly pick up 70% of English Web pages from each class as labeled data. On the other hand, because in real world applications the Web pages in the target language are mostly unlabeled, we simulate two different cases: (1) no labeled Web pages in the target languages and (2) we randomly pick up 20% Web pages from each class as labeled data in the concerned target language. As a result, we end up with six testing data sets, which are summarized in Table III. For each testing data set, our task is to classify the unlabeled Web pages in the corresponding target language.

6.2. Evaluation Metrics

Two widely used classification performance metrics in statistical learning and IR are used in our experiments: macro-average precision and F_1-measure. Let f be the function which maps from document d to its true class label $c = f(d)$, and h be the function which maps from document d to its prediction label $c = h(d)$ given by the classifiers. The macro-average precision P and recall R are defined as:

$$P = \frac{1}{\mathcal{C}} \sum_{c \in \mathcal{C}} \frac{\{d | d \in X_c \wedge h(d) = f(d) = c\}}{\{d | d \in X_c \wedge h(d) = c\}} \tag{37}$$

$$R = \frac{1}{\mathcal{C}} \sum_{c \in \mathcal{C}} \frac{\{d | d \in X_c \wedge h(d) = f(d) = c\}}{\{d | d \in X_c \wedge f(d) = c\}} \tag{38}$$



(a) α. (b) β. (c) γ. (d) μ.

Fig. 3. Classification performance (measured by macro-average precision) with respect to different parameter settings of the proposed DKT approach on $D2$ data set, which show that our approach is stable with respect to a wide range of parameters settings.

The F_1 measure is the harmonic mean of precision and recall, which is defined as follows:

$$F_1 = \frac{2PR}{P+R} \tag{39}$$

6.3. Stopping Criterion of Our Iterative Algorithms

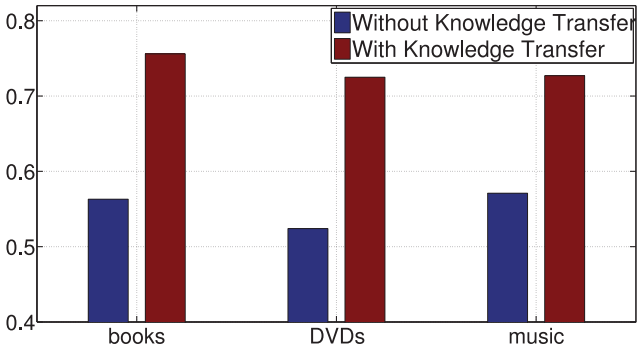
In order to solve the two proposed objectives in Equations (15) and (28), we presented two iterative algorithms in Algorithm 1 and Algorithm 2, which employ the following stopping criterion:

$$\frac{\left\|\mathbf{Z}^{(k+1)} - \mathbf{Z}^{(k)}\right\|}{\max\left(\left\|\mathbf{Z}^{(k)}\right\|, 1\right)} \leq \text{Tol}, \tag{40}$$

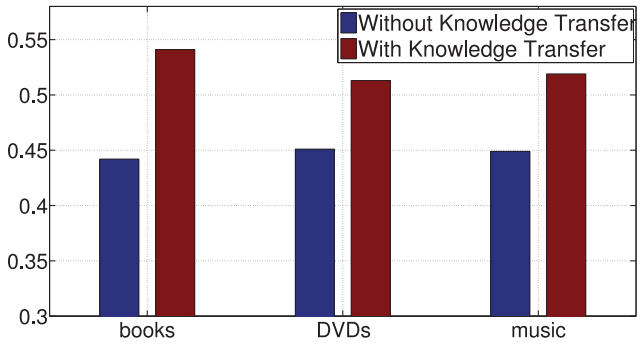
where $\mathbf{Z}$ is a concerned variable and Tol is a small number. In our experiments, we set $\text{Tol} = 10^{-4}$. When all the variables of an running algorithm, that is, $\mathbf{F}^a$, $\mathbf{F}^t$, $\mathbf{S}^a$, $\mathbf{S}^t$, $\mathbf{G}^a$, or $\mathbf{G}^t$, satisfy the criterion in Equation (40), we stop the iteration procedures.

6.4. Study on Parameters

Because the proposed DKT and F-DKT approaches have four parameters, that is, α, β, γ and μ in Equations (15) and (28), we first evaluate their impacts on the classification performance. Although it is tedious to seek an optimal combination of them, we can demonstrate that the performance of our DKT and F-DKT approaches are not sensitive when the parameters are sampled in some value ranges. On our preliminary tests, we bound the parameters in the ranges of $1 \leq \alpha \leq 10$, $1 \leq \beta \leq 10$, $0.5 \leq \gamma \leq 3$ and $1 \leq \mu \leq 5$. We report classification performance on test data set $D2$ in Figure 3 as it has labeled Web pages in both auxiliary language and target language. In our experiments, the default parameter values are set as $\alpha = 1$, $\beta = 4.5$, $\gamma = 1$, $\mu = 1$, that is, when evaluating one parameter, we set the other three parameters to the default



(a) Precision.



(b) F_1 score.

Fig. 4. Comparison of the proposed approach (Equation (15)) and its degenerate version (Equation (41)) on $D1$ test data.

values as listed. For each set of parameter combination, we repeat the experiment for 50 times and the average performance is reported as one point in Figure 3.

From Figure 3, we can see that the average performances of our DKT approach with respect to each parameter remain considerable stable in a large range of the parameter settings. This demonstrates that our approach is very robust again parameter settings and thereby suitable for practical use.

6.5. Effectiveness of Knowledge Transfer in Cross-Language Web Page Classification

Because the main purpose of the proposed DKT approach is to transfer knowledge from the Web pages in an auxiliary language to those in another target language which does not have sufficient labeled data, we first evaluate the knowledge transfer capability of the proposed approach in this subsection. We compare the proposed DKT approach against its degenerate version as:

$$\begin{aligned} \min\ J &= \|\mathbf{X}^t - \mathbf{F}^t\mathbf{S}^t\left(\mathbf{G}^t\right)^T\|^2 \\ &\quad + \mathbf{tr}[\beta(\mathbf{G}^t - \mathbf{Y}^t)^T\mathbf{C}^t(\mathbf{G}^t - \mathbf{Y}^t)], \\ &\quad s.t.\quad \mathbf{S}^t \geq 0, \mathbf{F}^t \geq 0, \mathbf{G}^t \geq 0, (\mathbf{F}^t)^T\mathbf{F}^t = \mathbf{I}, (\mathbf{G}^t)^T\mathbf{G}^t = \mathbf{I}, \end{aligned} \tag{41}$$

in which knowledge transfer terms in Equation (15) are removed. Thus Equation (41) is a semi-supervised learning method working with the target data only, whose similar form was ever proposed in Gu and Zhou [2009b].

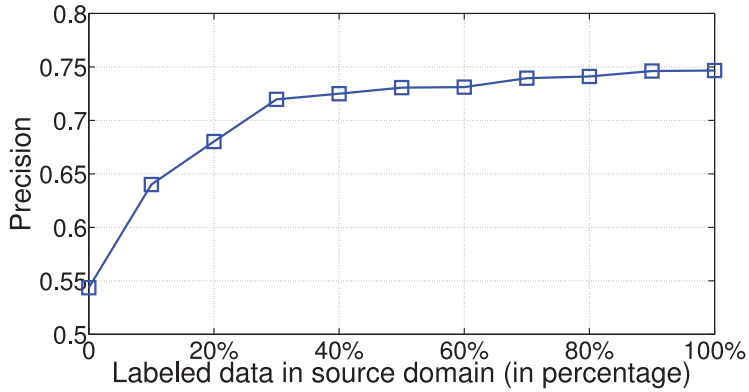


Fig. 5. Average precisions over all the classes on the target data set when the amount of labeled Web pages in the auxiliary language varies.

We conduct this experiment on $D2$ data set. According to the experimental results in Section 6.4, we set the tradeoff parameters $\alpha = 1$, $\beta = 4.5$, $\gamma = 1$ and $\mu = 1$, following [Ding et al. 2006; Gu and Zhou 2009a] the number of word clusters is set to same as Web page classes $k_1 = k_2 = 3$. We predict labels for the 80% unlabeled Web pages in German and repeat the experiments for 100 times. The average classification performances measured by precision and F_1 score for each class are shown in Figure 4(a) and Figure 4(b) respectively, from which we can see that our approach using knowledge transfer outperforms its degenerate version without using knowledge transfer in all the three classes. These observations concretely demonstrate the usefulness of knowledge transfer in cross-language Web page classification.

In order to evaluate the detailed impact of training information in the auxiliary data to the classification performance on the target data, we vary the amount of labeled Web pages in the auxiliary language and examine the corresponding classification performance on the Web pages in the target language of our approach. The average precisions over all the three classes for different amount of labeled Web pages in the auxiliary language are reported in Figure 5, which show that the more labeled data we have in the auxiliary domain, the better classification performance we can achieve on the target data set. This is consistent with the theoretical analysis, and again confirms the effectiveness of our approach to transfer knowledge in cross-language Web page classification.

6.6. Comparisons to Related Methods

Now we evaluate the proposed DKT approach by comparing it to two most recent transfer learning methods including (1) KTW method [Li et al. 2009], (2) Matrix Trifactorization based classification framework (MTrick) [Zhuang et al. 2010], and (3) a very recent cross-language Web classification method using IB theory [Ling et al. 2008]. These methods have demonstrated state-of-the-art classification performance in a variety of real world applications.

We also compare the proposed methods in the current manuscript to that in our previous conference publication [Wang et al. 2011b]. Because the latter transfers the supervised knowledge via fixed middle factor matrix $\mathbf{S}$ in the data matrix factorizations, we denote it as "DKT (S fixed)" in Tables (IV–VI).

In addition, we also report the classification performances of Support Vector Machine (SVM) (supervised method), and Transductive SVM (TSVM) [Joa] (semi-supervised method) as baselines.

Table IV. Macro-Average Precision and F_1 Measure of Compared Methods on English-German Web Page Data Sets

	*D*1		*D*2	
Methods	Precision	F1	Precision	F1
SVM_T	–	–	0.679	0.468
SVM_TS	0.682	0.479	0.682	0.481
TSVM_T	–	–	0.682	0.475
TSVM_TS	0.689	0.483	0.701	0.489
KTW	0.673	0.481	0.675	0.483
MTrick	0.695	0.490	0.699	0.492
IB	0.691	0.492	0.703	0.501
DKT (S only)	0.697	0.495	0.718	0.505
DKT (S fixed)	0.716	0.508	0.730	0.510
DKT	0.735	0.517	0.748	0.532
F-DKT	0.721	0.513	0.734	0.527

Table V. Macro-Average Precision and F_1 Measure of Compared Methods on English-French Web Page Data Sets

	*D*3		*D*4	
Methods	Precision	F1	Precision	F1
SVM_T	–	–	0.663	0.452
SVM_TS	0.670	0.470	0.628	0.469
TSVM_T	–	–	0.670	0.461
TSVM_TS	0.675	0.475	0.687	0.472
KTW	0.663	0.470	0.669	0.471
MTrick	0.682	0.481	0.681	0.481
IB	0.680	0.480	0.690	0.486
DKT (S only)	0.683	0.483	0.702	0.492
DKT (S fixed)	0.701	0.498	0.718	0.501
DKT	0.721	0.511	0.730	0.515
F-DKT	0.713	0.504	0.723	0.506

Table VI. Macro-Average Precision and F_1 Measure of Compared Methods on English–Japanese Web Page Data Sets

	*D*3		*D*4	
Methods	Precision	F1	Precision	F1
SVM_T	–	–	0.651	0.447
SVM_TS	0.662	0.463	0.676	0.460
TSVM_T	–	–	0.663	0.456
TSVM_TS	0.668	0.468	0.676	0.467
KTW	0.656	0.462	0.658	0.463
MTrick	0.674	0.472	0.672	0.475
IB	0.672	0.470	0.681	0.478
DKT (S only)	0.679	0.477	0.695	0.486
DKT (S fixed)	0.688	0.485	0.707	0.493
DKT	0.701	0.497	0.719	0.512
F-DKT	0.690	0.491	0.713	0.503

6.6.1. Experimental Setups. SVM and TSVM methods can use either the labeled data in the target language or the labeled data in both the auxiliary and target languages. We refer to SVM_T, TSVM_T as the former case, and SVM_ST, TSVM_ST as the latter case. For the latter case, the data from the both auxiliary and target languages are used in a homogeneous way. This is equivalent to assume the Web pages from different laguages are drawn from a same distribution, which, however, is not true in reality. Following previous works, for the both methods, we train one-versus-others

classifiers, with the fixed regularization parameter $C = 1$. Gaussian kernel is used (*i.e.*, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$) where γ is set as $1/m$. SVM and TSVM are implemented by SVM$^{\text{light}}$ [Joachims 2008].

The parameters of KTW and MTrick are set as optimal following their original works [Li et al. 2009; Zhuang et al. 2010]. The iteration number of IB method is set as 100.

For our approach, we follow the same settings as in Section 6.5. Due to the nature of our optimization objective in Equation (15), we always use S, that is, the associations between word clusters and Web page classes, to transfer knowledge. In order to test the flexibility of our approach, we consider two different cases of our approach for using words to transfer knowledge: (1) not use words transfer denoted as "DKT (S only)", that is, set $\gamma = 0$ in Equation (15); and (2) use words transfer denoted as "DKT".

6.6.2. Experimental Results. Tables (IV–VI) present the classification performances measured by macro-average precision and F_1 score of the compared methods on six different test data sets. A number of interesting observations can be seen from these results.

First, the proposed DKT approach consistently outperforms the other compared methods, sometimes very significantly. DKT (S only) method is always worse than DKT approach, which confirms the usefulness of the knowledge transfer path by words. In addition, the classification performances of the DKT (S fixed) method proposed in our previous conference publication [Wang et al. 2011b] is also not as good as the DKT approach, which demonstrate that the new objective in Equation (13) is more effective to transfer supervised knowledge than Equation (14) proposed in Wang et al. [2011b], which is a new contribution of the current manuscript compared to our previous work in Wang et al. [2011b].

Second, although the classification performance of F-DKT is satisfactory, which is not as good as DKT method. This is consistent with their formulations in that the former is more stringently constrained to use cluster indicator matrices as factor matrices and thereby has less data representation power. However, as demonstrated shortly in Section 6.7, the computational speed of F-DKT method is much faster than that of DKT method, which makes it more suitable for practical use.

Third, from the experimental results of SVM_ST and TSVM_ST methods, we can see that considering Web pages from different languages as homogenous typically leads to unsatisfactory classification performance. Because the cross-domain methods, including ours, are generally better than these two methods, knowledge transfer from the auxiliary language to the target one is important to improve the classification performance.

Fourth, our DKT approach is able to transfer knowledge in two ways, that is, words and the associations between word clusters and Web page classes. Therefore it achieves encouraging classification performance on all the six test data sets. In contrast, KTW method can only transfer knowledge through words, and MTrick method only transfers knowledge through the associations between word clusters and Web page classes, their performances are generally not as good as other transfer learning methods.

Last, but not the least, our approach is able to exploit the label information in both auxiliary and target data, whereas KTW method and MTrick method cannot benefit from label information in target domain, and SVM_T method and TSVM_T method cannot work with label information in auxiliary data. The more labeled data in target domain, the better classification performance our approach can achieve. In summary, all the earlier observations demonstrate the effectiveness of the proposed DKT approach in cross-language Web page classification.

6.7. Studies of Computational Speeds

Because one of the main contributions of this article is to present a fast version of the proposed DKT approach, in this section we evaluate the computation efficiencies

Table VII. Run Time (in seconds) of the Proposed DKT Method and F-DKT Method

Data set	$D1$	$D2$	$D3$	$D4$	$D5$	$D6$
DKT	7.54×10^4	7.25×10^4	7.81×10^4	7.36×10^4	7.47×10^4	7.91×10^4
F-DKT	8.96×10^3	8.26×10^3	9.15×10^3	8.69×10^3	8.83×10^3	8.93×10^3

Table VIII. Run Time (in seconds) on the Three Simulated Large-Scale Data

				Run time (sec)	
Data set	# Web pages	# keywords	# classes	DKT	F-DKT
$S1$	20,000	1,000	20	7.93×10^5	9.87×10^4
$S2$	200,000	1,000	20	1.64×10^7	2.15×10^6
$S3$	2,000,000	1,000	20	3.84×10^9	6.79×10^7

of the presented F-DKT approach, and compare it with its continuous version. All our experiments are performed on a Dell PowerEdge 2900 server, which has two quad-core Intel Xeon 5300 sequence CPU processors at 3.0 GHz and 48G bytes memory.

6.7.1. Experiments on Real Cross-Language Web Page Classification Data. We first evaluate the computational efficiencies of the proposed DKT method and F-DKT method by comparing their running time on the six test data sets. We repeat each experiment for 10 times and report the average run time in Table VII. From the results in Table VII we can see that, as expected, the F-DKT method is much faster than DKT method, which demonstrate the usefulness of introducing new constraints on the factor matrices to be cluster indicator matrices.

6.7.2. Experiments on Simulated Large-Scale Data. In order to further evaluate the computational efficiency of the proposed F-DKT approach, we perform experiments on three simulated large-scale data. In the our experiments, we randomly create two data matrices to simulate the two input data sets in the auxiliary and target language respectively. We simulate three different conditions in which the numbers of Web pages are 20,000, 200,000 and 2,000,000 respectively. We assume the number of the keywords of all the data sets is 1,000. As a results, we end up with three simulated data sets, $S1$, $S2$ and $S3$, which are summarized in Table VIII. For all the simulated test data, we assume to have 20 semantic classes. We randomly pick up 70% of the auxiliary data points and randomly label them to one of the 20 classes. We also randomly pick up 20% of the target data and randomly label them to one of the 20 classes. Then we run our DKT approach by solving Equation (15) using Algorithm 1, and F-DKT approach by solving Equation (28) using Algorithm 2. We repeat each experiment for 20 times and report the average run time in Table VIII. Again, the results clearly demonstrate the computational advantage of the F-DKT approach over DKT approach, which provide one evidence to support the usefulness of using cluster indicator matrices in NMTF.

7. CONCLUSIONS

In this article, we proposed a novel joint NMTF based DKT approach for cross-language Web page classification. Our approach adopts the idea of transfer learning to pass knowledge across languages but not simply combine the Web page data from different languages. By carefully examine the cross-language Web page classification problem, we observe that common semantic patterns usually exist in Web pages for a same topic from different languages. Moreover, we also observe that the associations between word clusters and Web page classes are more reliable to transfer knowledge than using raw words. With these recognitions, our approach is designed to transfer knowledge across languages through two different ways: word clusters and the associations between word clusters and Web pages classes. With this enhanced knowledge transfer, our approach is able to address the main challenges in cross-language Web page classification: cultural discrepancies, translation ambiguities and data diversity. In order to

deal with large-scale real world Web page data sets, we further develop the proposed DKT method by constraining the factor matrices to cluster indicator matrices, a special type of nonnegative matrices. Due to the nature of the cluster indicator matrices, the optimization problem of the proposed approach is decoupled, which leads to subproblems in much smaller sizes involving much less matrix multiplications. As a result, our F-DKT approach is much more computationally efficient, though it uses a more stringent constraints than traditional nonnegative constraints. Extensive experiments using a real world Web page data set demonstrated encouraging results from a number of aspects that validate our approach.

APPENDIX: ANALYSIS OF ALGORITHM CONVERGENCE

In this section, we will investigate the convergence of Algorithm 1. We use the auxiliary function approach [Lee and Seung 2001] to prove the convergence of the algorithm.

LEMMA A.1 [LEE AND SEUNG 2001]. *$Z(h, h')$ is an auxiliary function of $F(h)$ if the conditions $Z(h, h') \geq F(h)$ and $Z(h, h') = F(h)$ are satisfied.*

LEMMA A.2 [LEE AND SEUNG 2001]. *If Z is an auxiliary function for F, then F is non-increasing under the update $h^{(t+1)} = \arg\min_h Z(h, h')$.*

LEMMA A.3 [DING ET AL. 2006]. *For any matrices $\mathbf{A} \in \Re_+^{n\times n}$, $\mathbf{B} \in \Re_+^{k\times k}$, $\mathbf{S} \in \Re_+^{n\times k}$ and $\mathbf{S}' \in \Re_+^{n\times k}$, and $\mathbf{A}$ and $\mathbf{B}$ are symmetric, the following inequality holds*

$$\sum_{ip} \frac{(\mathbf{AS}'\mathbf{B})_{ip}\,\mathbf{S}_{ip}^2}{\mathbf{S}'_{ip}} \geq \mathbf{tr}(\mathbf{S}^T\mathbf{ASB}). \tag{42}$$

THEOREM A.4. *The Lagrangian function L in Equation (18) can be written as*

$$\begin{aligned} L(\mathbf{F}^a) = \mathbf{tr}[&-2(\mathbf{X}^a)^T\mathbf{F}^a\mathbf{S}^a(\mathbf{G}^a)^T + (\mathbf{S}^a(\mathbf{G}^a)^T\mathbf{G}^a(\mathbf{S}^a)^T + \mathbf{U})(\mathbf{F}^a)^T\mathbf{F}^a \\ &- 2\gamma(\mathbf{F}^t)^T\mathbf{VF}^a + \gamma(\mathbf{F}^a)^T\mathbf{VF}^a], \end{aligned} \tag{43}$$

in which the constant terms are removed. Then the following function

$$\begin{aligned} H(\mathbf{F}^a, \mathbf{F}^{a\prime}) = &-2\sum_{ij}((\mathbf{X}^a)^T\mathbf{F}^a\mathbf{S}^a(\mathbf{G}^a)^T) \\ &+ \sum_{ij}[\mathbf{F}^{a\prime}(\mathbf{S}^a(\mathbf{G}^a)^T\mathbf{G}^a(\mathbf{S}^a)^T + \mathbf{U})]\frac{\left(\mathbf{F}^a_{(ij)}\right)^2}{\mathbf{F}^{a\ \prime}_{(ij)}} \\ &- 2\gamma\sum_{ij}(\mathbf{VF}^t)_{(ij)}\mathbf{F}^{a\ \prime}_{(ij)}\left(1 + \log\frac{\mathbf{F}^a_{(ij)}}{\mathbf{F}^{a\ \prime}_{(ij)}}\right) \\ &+ \gamma\sum_{ij}(\mathbf{VF}^{a\prime})_{(ij)}\frac{\left(\mathbf{F}^a_{(ij)}\right)^2}{\mathbf{F}^{a\ \prime}_{(ij)}} \end{aligned} \tag{44}$$

is an auxiliary function for $L(\mathbf{F}^a)$. Furthermore, it is a convex function in $\mathbf{F}^a$ and its global minimum is

$$\mathbf{F}^a_{(ij)} = \mathbf{F}^a_{(ij)}\sqrt{\frac{(\mathbf{X}^a\mathbf{G}^a(\mathbf{S}^a)^T + \gamma\mathbf{VF}^t)_{(ij)}}{(\mathbf{F}^a(\mathbf{F}^a)^T\mathbf{X}^a\mathbf{G}^a(\mathbf{S}^a)^T + \gamma\mathbf{VF}^a)_{(ij)}}}. \tag{45}$$

PROOF. According to Lemma A.3, we have

$$\mathbf{tr}((\mathbf{S}^a(\mathbf{G}^a)^T\mathbf{G}^a(\mathbf{S}^a)^T+\mathbf{U})(\mathbf{F}^a)^T\mathbf{F}^a) \le \sum_{ij}[\mathbf{F}^{a\prime}(\mathbf{S}^a(\mathbf{G}^a)^T\mathbf{G}^a(\mathbf{S}^a)^T+\mathbf{U})]\frac{\left(\mathbf{F}^a_{(ij)}\right)^2}{\mathbf{F}^{a\ \prime}_{(ij)}}, \tag{46}$$

$$\mathbf{tr}((\mathbf{F}^a)^T\mathbf{V}\mathbf{F}^a) \le \sum_{ij}(\mathbf{V}\mathbf{F}^{a\prime})_{(ij)}\frac{\left(\mathbf{F}^a_{(ij)}\right)^2}{\mathbf{F}^{a\ \prime}_{(ij)}}. \tag{47}$$

Because $z \le 1+\log\ z,\ \forall\, z>0$, we have

$$\mathbf{tr}((\mathbf{F}^t)^T\mathbf{V}\mathbf{F}^a) \ge \sum_{ij}(\mathbf{V}\mathbf{F}^t)_{(ij)}\mathbf{F}^{a\ \prime}_{(ij)}\left(1+\log\frac{\mathbf{F}^a_{(ij)}}{\mathbf{F}^{a\ \prime}_{(ij)}}\right). \tag{48}$$

Summing over all the bounds in Equations (46–48), we can obtain $H(\mathbf{F}^a,\mathbf{F}^{a\prime})$, which clearly satisfies (1) $H(\mathbf{F}^a,\mathbf{F}^{a\prime}) \ge J(\mathbf{F}^a)$ and (2) $H(\mathbf{F}^a,\mathbf{F}^a) = J(\mathbf{F}^a)$.

Then, fixing $\mathbf{F}^{a\prime}$, we minimize $H(\mathbf{F}^a,\mathbf{F}^{a\prime})$.

$$\begin{aligned}\frac{\partial H(\mathbf{F}^a,\mathbf{F}^{a\prime})}{\partial \mathbf{F}^a_{(ij)}} =& -2\left[(\mathbf{X}^a\mathbf{G}^a(\mathbf{S}^a)^T)_{(ij)}+\gamma(\mathbf{V}\mathbf{F}^t)_{(ij)}\right]\frac{\mathbf{F}^{a\ \prime}_{(ij)}}{\mathbf{F}^a_{(ij)}}\\ &+2\left[(\mathbf{F}^{a\prime}(\mathbf{S}^a(\mathbf{G}^a)^T\mathbf{G}^a(\mathbf{S}^a)^T+\mathbf{U}))_{(ij)}+\gamma(\mathbf{V}\mathbf{F}^{a\prime})_{(ij)}\right]\frac{\mathbf{F}^a_{(ij)}}{\mathbf{F}^{a\ \prime}_{(ij)}}\end{aligned} \tag{49}$$

and the Hessian matrix of $H(\mathbf{F}^a,\mathbf{F}^{a\prime})$ is

$$\begin{aligned}\frac{\partial^2 H(\mathbf{F}^a,\mathbf{F}^{a\prime})}{\mathbf{F}^a_{(ij)}\mathbf{F}^a_{(kl)}} =& \delta_{ik}\delta_{jl}\left\{2\left[(\mathbf{X}^a\mathbf{G}^a(\mathbf{S}^a)^T)_{(ij)}+\gamma(\mathbf{V}\mathbf{F}^t)_{(ij)}\right]\frac{\mathbf{F}^{a\ \prime}_{(ij)}}{\left(\mathbf{F}^a_{(ij)}\right)^2}\right.\\ &\left.+2\left[(\mathbf{F}^{a\prime}(\mathbf{S}^a(\mathbf{G}^a)^T\mathbf{G}^a(\mathbf{S}^a)^T+\mathbf{U}))_{(ij)}+\gamma(\mathbf{V}\mathbf{F}^{a\prime})_{(ij)}\right]\mathbf{F}^{a\ \prime}_{(ij)}\right\},\end{aligned} \tag{50}$$

which is a diagonal matrix with positive diagonal elements. Therefore $H(\mathbf{F}^a,\mathbf{F}^{a\prime})$ is a convex function of $\mathbf{F}^a$, and we can obtain the global minimum of $H(\mathbf{F}^a,\mathbf{F}^{a\prime})$ by setting $\partial H(\mathbf{F}^a,\mathbf{F}^{a\prime})/\partial\mathbf{F}^a_{(ij)} = 0$ and solving for $\mathbf{F}^a$, from which we get Equation (45). This completes the proof of Theorem A.4. □

THEOREM A.5. *Updating $\mathbf{F}^a$ using the rule in Algorithm 1 will monotonically decreases the value of the objective $J(\mathbf{F}^a)$ in Equation (43), thus it finally converge.*

PROOF. By Lemma A.1 and Theorem A.4, we can get that $J[(\mathbf{F}^a)^0] = H[(\mathbf{F}^a)^0,(\mathbf{F}^a)^0] \ge H[(\mathbf{F}^a)^1,(\mathbf{F}^a)^0] \ge J[(\mathbf{F}^a)^1]\ldots$ So $J(\mathbf{F}^a)$ is monotonically decreasing. As $J(\mathbf{F}^a$ is clearly bounded later, we prove this theorem. □

THEOREM A.6. *Updating $\mathbf{G}^a$, $\mathbf{S}^a$, $\mathbf{S}^t$, $\mathbf{F}^t$, and $\mathbf{G}^t$ using the rules in Algorithm 1, the respective objective will converge.*

Theorem A.6 can be similarly proved as Theorems (A.4–A.5).

Because J in Equation (15) is obviously lower bounded by 0, Algorithm 1 is guaranteed to converge by Theorems (A.5–A.6).

REFERENCES

Nuria Bel, Cornelis H. A. Koster, and Marta Villegas. 2003. Cross-lingual text categorization. *Research and Advanced Technology for Digital Libraries.* Lecture Notes in Computer Science, Vol. 2769. 126–139.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 120–128.

Gang Chen, Fei Wang, and Changshui Zhang. 2009. Collaborative filtering using orthogonal nonnegative matrix tri-factorization. *Information Processing and Management* 45, 3 (2009), 368–379.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 210–219.

Chris Ding and Xiaofeng He. 2004. K-means clustering via principal component analysis. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*. ACM, New York, NY, USA, 29.

Chris Ding, Xiaofeng He, and Horst D. Simon. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, Philadelphia, PA, 606–610.

Chris H. Q. Ding, Tao Li and Michael. I. Jordan. 2010. Convex and semi-nonnegative matrix factorizations. *TPAMI* 32, 1 (2010), 45–55.

Chris Ding, Tao Li, Wei Peng, and Haesun Park. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 126–135.

Quanquan Gu and Jie Zhou. 2009a. Co-clustering on manifolds. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 359–368.

Quanquan Gu and Jie Zhou. 2009b. Transductive classification via dual regularization. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*. Springer-Verlag, Berlin, Heidelberg, 439–454.

Quanquan Gu, Jie Zhou, and Chris Ding. 2010. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In *Proceedings of the 10th SIAM International Conference on Data Mining*. SIAM, Philadelphia, PA, 199–210.

Thorsten Joachims. 2008. SVMLight: Support vector machine. http://svmlight.joachims.org/.

Daniel D. Lee and Hyunjune S. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

Daniel D. Lee and Hyunjune S. Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, USA, 556–562.

Tao Li, Vikas Sindhwani, Chris Ding, and Yi Zhang. 2009. Knowledge transformation for cross-domain sentiment classification. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 716–717.

Tao Li, V. Sindhwani, C. Ding, and Y. Zhang. 2010. Bridging domains with words: Opinion analysis with matrix tri-factorizations. In *SDM*.

Xiao Ling, Gui-Rong Xue, Wenyuan Dai, Yun Jiang, Qiang Yang, and Yong Yu. 2008. Can Chinese web pages be classified with English data source? In *Proceedings of the 17th International Conference on World Wide Web*. ACM, New York, NY, USA, 969–978.

J. S. Olsson, Douglas W. Oard, and Jan Hajič. 2005. Cross-language text classification. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 645–646.

Sinno J. Pan and Qiang Yang. 2009. A survey on transfer learning. In *IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE)* 22, 10, 1345–1359.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1118–1127.

Gabriela Ramírez-de-la-Rosa, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, David Pinto-Avendaño, Thamar Solorio. 2010. Using information from the target language to improve crosslingual text classification. In *Proceedings of 7th International Conference on NLP, IceTAL 2010*. Reykjavik, Iceland, 305–313.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1057–1067.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1–Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, 235–243.

Fei Wang, Tao Li, and Changshui Zhang. 2008. Semi-supervised clustering via matrix factorization. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, Philadelphia, PA, 1–12.

Hua Wang, Feiping Nie, Heng Huang, and Fillia Makedon. 2011a. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume 2*. AAAI, USA, 1553–1558.

Hua Wang, Heng Huang, Feiping Nie, and Chris Ding. 2011b. Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 933–942.

Hua Wang, Heng Huang, and Chris Ding. 2011c. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, New York, NY, USA, 279–284.

Hua Wang, Feiping Nie, Heng Huang, and Chris Ding. 2011d. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. In *IEEE 11th International Conference on Data Mining (ICDM)*. IEEE, Vancouver, BC, 774–783.

Hua Wang, Feiping Nie, Heng Huang, and Chris Ding. 2011e. Dyadic transfer learning for cross-domain image classification. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, Barcelona, Spain, 551–556.

Ke Wu and Bao-Liang Lu. 2008. A refinement framework for cross language text categorization. In *Proceedings of the 4th Asia Information Retrieval Conference on Information Retrieval Technology*. Lecture Notes in Computer Science, Vol. 4993. Springer-Verlag, Harbin, China, 401–411.

Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. 2001. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, USA, 1057–1064.

Fuzhen Zhuang, Ping Luo, Hui Xiong, Qing He, Yuhong Xiong, and Zhongzhi Shi. 2010. Exploiting associations between word clusters and document classes for cross-domain text categorization. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, Philadelphia, PA, 13–24.

Received July 2013; revised May 2014; accepted December 2014