Cross-Language Web Page Classification via Dual Knowledge Transfer Using Nonnegative Matrix Tri-Factorization

Hua Wang
Department of Computer
Science and Engineering
University of Texas at
Arlington, TX 76019, USA
huawangcs@gmail.com

Heng Huang
Department of Computer
Science and Engineering
University of Texas at
Arlington, TX 76019, USA
heng@uta.edu

Feiping Nie
Department of Computer
Science and Engineering
University of Texas at
Arlington, TX 76019, USA
feipingnie@gmail.com

Chris Ding
Department of Computer
Science and Engineering
University of Texas at
Arlington, TX 76019, USA
chqding@uta.edu

ABSTRACT

The lack of sufficient labeled Web pages in many languages, especially for those uncommonly used ones, presents a great challenge to traditional supervised classification methods to achieve satisfactory Web page classification performance. To address this, we propose a novel Nonnegative Matrix Tri-factorization (NMTF) based Dual Knowledge Transfer (DKT) approach for cross-language Web page classification, which is based on the following two important observations. First, we observe that Web pages for a same topic from different languages usually share some common semantic patterns, though in different representation forms. Second, we also observe that the associations between word clusters and Web page classes are a more reliable carrier than raw words to transfer knowledge across languages. With these recognitions, we attempt to transfer knowledge from the auxiliary language, in which abundant labeled Web pages are available, to target languages, in which we want classify Web pages, through two different paths: word cluster approximations and the associations between word clusters and Web page classes. Due to the reinforcement between these two different knowledge transfer paths, our approach can achieve better classification accuracy. We evaluate the proposed approach in extensive experiments using a real world cross-language Web page data set. Promising results demonstrate the effectiveness of our approach that is consistent with our theoretical analyses.

SIGIR'11, July 24–28, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0757-4/11/07 ...$10.00.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Cross-language Classification, Knowledge Transfer, Nonnegative Matrix Factorization

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 April 2011, over 131.1 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 becomes 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 the lack of sufficient labeled data. This prevents us from training an effective classification model, which usually requires a large amount of labeled data. Statistically speaking, the more labeled training data one can obtain, the more accurate and robust the classification model is. Fortunately, due to many reasons, there exists a lot of labeled Web pages in several most commonly used languages, such as English. It is hence useful and intriguing to make use of these labeled Web pages in one language, called as *auxiliary language*, to help to classify Web pages in another language, called as *target language*. This problem is called as *cross-language Web page classification* [15]. In this paper, we explore this important, yet challenging, problem by proposing a novel

[1] `http://www.domaintools.com/internet-statistics/`

Nonnegative Matrix Tri-factorization (NMTF) based Dual Knowledge Transfer (DKT) approach.

1.1 Challenges in Cross-Language Web Page Classification and Our Motivations

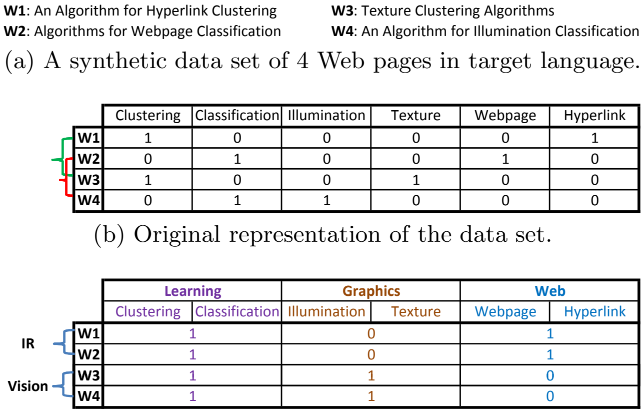
One of the most widely used strategy in cross-language Web page (text) classification is using language translation [15, 16, 18–21, 23]. One can either translate test data into the auxiliary language, or translate training data into the target language, and then train and classify the resulted data in one single language. Although this straightforward method may be feasible, it suffers from a number of critical problems that impede its practical use [15, 19, 20]. In this subsection, we examine the challenges in cross-language Web page classification and seek opportunities to overcome them, which motivate our approach.

Cultural discrepancy. The first difficulty in cross-language Web page classification is due to cultural discrepancies, which heavily impact the classification performance in spite of a perfect translation [15, 19]. Given that a language is the way of expression of a cultural and socially homogenous community, Web pages from a same category but different languages may concern very different topics. For example, we consider Web pages that report sports news in France (in French) and in USA (in English). While the former typically pays more attention to soccer, rugby and cricket, the latter is more interested in basketball and American football. From machine learning perspective of view, this corresponds to the situation where 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.

Moreover, 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 [16], the English speakers tend to contribute more to some topics than their Czech counterparts (*e.g.*, to discuss "London" more than "Prague"), so that, having only data in English, we may expect them to do poorly at identifying topics like "Prague". Czech speakers, on the other hand, often talk about "Prague", so that by leveraging Czech data, we may expect to improve on detecting topic "Prague" in English Web pages.

To overcome this problem, instead of simply combining the data, we consider to transfer labeling information contained in Web pages in the auxiliary language to those in the target language [17]. Our approach is based on the observation that Web pages in different languages from a same category often share the same semantic information, although they are in different representation forms, *e.g.*, French words and English words [15]. Therefore, we may abstract the prior knowledge in the auxiliary language into semantic patterns, and make use of them to help to classify Web pages in the target language. To transfer knowledge across languages, the most natural carrier is the basic linguistic representation unit — words. We give an example to illustrate the usefulness of knowledge transfer by words in Web page classifications as in Figure 1.

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 clusters them (the rows of the data matrix) based on 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 with 3 semantic 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. This is a very meaningful result, because the former is concerned with "information retrieval", while the latter is interested in "vision". More theoretical analysis for this example will be given later in Section 3.2.



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) Original representation of the data set.

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

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

Figure 1: An illustrative example to demonstrate the usefulness of leveraging the prior knowledge learned from an auxiliary language when clustering Web pages in a target language.

In our approach, the first path to transfer knowledge across languages is by word cluster approximations, which is schematically shown by the red lines in Figure 2.

Translation ambiguity. In the process of language translation, the ambiguities introduced by dictionaries is another important challenge in cross-language Web page classification. For example, the word "阅读材料 (reading materials)" in Chinese Web pages could be reasonably translated as "textbooks", "required reading list", "reference" and so on. Since the linguistic habits in expressing a concept are different in different languages, the phrases for a same concept may have different probabilities in different languages. Therefore, transferring knowledge by the raw words sometimes are not reliable. However, the concept behind the phrases may have the same effect to indicate the 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". In other words, only the concept behind raw words are stable in indicating taxonomy, and the association between word clusters and Web page categories is independent of languages [25]. Therefore, we use it as the second bridge to transfer knowl-

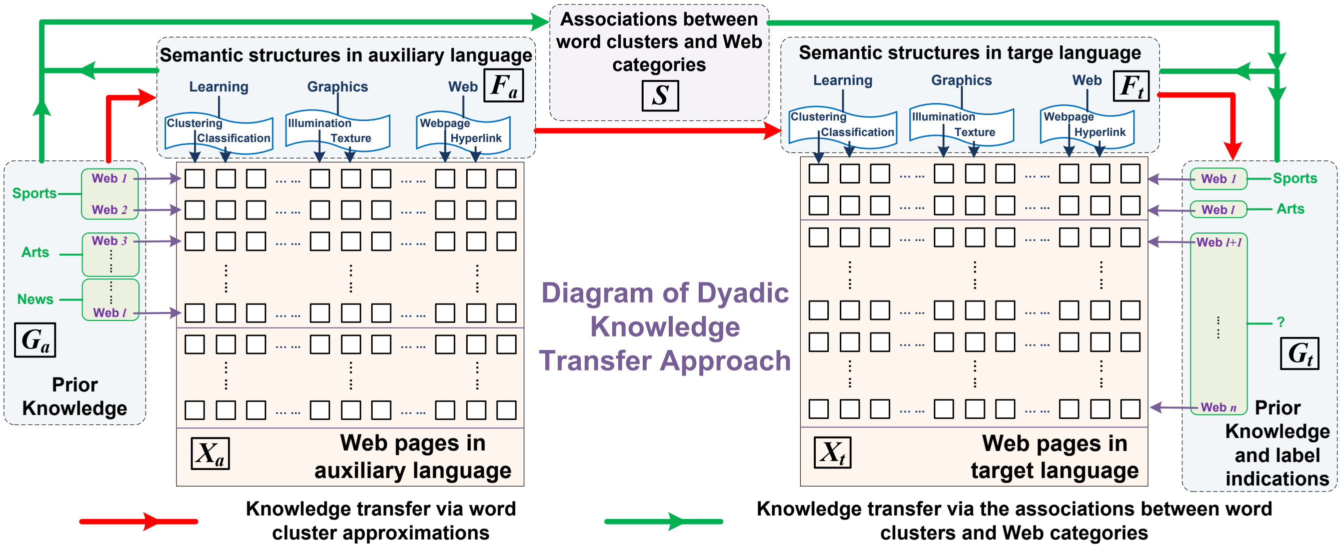

Associations between word clusters and Web categories
S
Semantic structures in auxiliary language
Learning
Graphics
Web
F_a
Clustering
Classification
Illumination
Texture
Webpage
Hyperlink
Semantic structures in targe language
Learning
Graphics
Web
F_t
Clustering
Classification
Illumination
Texture
Webpage
Hyperlink
Sports
Web 1
Web 2
Arts
Web 3
News
Web l
G_a
Prior Knowledge
Diagram of Dyadic Knowledge Transfer Approach
X_a
Web pages in auxiliary language
X_t
Web pages in target language
Web 1
Sports
Web l
Arts
Web l+1
?
Web n
G_t
Prior Knowledge and label indications
Knowledge transfer via word cluster approximations
Knowledge transfer via the associations between word clusters and Web categories

Figure 2: Diagram of the proposed Dual Knowledge Transfer (DKT) approach using NMTF. We transfer knowledge from auxiliary language to target language through two ways: *word cluster approximations* (F_a and F_t) and *the associations between word clusters and Web page categories* (S).

edge across different languages, which is illustrated by the green paths in Figure 2.

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 labeled by limited human effort. As a result, we can not rigidly assume the Web pages in the auxiliary language are always labeled and 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 tasks.

1.2 Our Contributions

Taking into account the three challenges in cross-language Web page classification, through a novel joint NMTF framework, we abstract the prior knowledge contained in Web pages in the auxiliary language, including both labeling information by human efforts and latent language structures, in two forms represented by the two factor matrices F_a and S of NMTF respectively, and then transfer them to the target language to guide the classification therein. The whole idea is summarized in Figure 2. Because we employ a two-way knowledge transfer, we call our proposed approach as *Dual Knowledge Transfer (DKT)* approach, which is interesting from a number of perspectives as following.

- In addressing the cross-language Web page classification problem, we observe the two possible paths to transfer knowledge across languages: the natural way by *word cluster approximations* and the reliable way by *associations between word clusters and Web page categories*. We propose a NMTF based DKT approach to make use of both of them.
- 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 where they are, either in auxiliary language or target language, or the both, our approach is always able to take advantage of the available labeling information.
- An efficient algorithm is presented to solve the proposed optimization objective, together with rigorous proof of its convergence.
- Extensive experiments on real world data sets demonstrate promising results that validate our approach.

2. A BRIEF REVIEW OF NMTF

In this section, we first briefly review NMTF and reveal how it transfers knowledge between data and features within a same data set, from which we will develop our approach.

Traditional Nonnegative Matrix Factorization (NMF) aims to find two nonnegative matrices whose product can well approximate the original nonnegative data matrix $X \in \mathbb{R}_+^{p\times n}$, *i.e.*, $X \approx FG^T$, where $F \in \mathbb{R}_+^{p\times k}$ and $G \in \mathbb{R}_+^{n\times k}$. The columns of X are data points and the rows of X are observations. An appropriate objective of NMF is to minimize [12]:

$$J_{\text{NMF}} = \|X - FG^T\|^2, \quad s.t. \quad F \geq 0, G \geq 0, \qquad (1)$$

where $\|\cdot\|$ denotes the Frobenius norm of a matrix. According to [5], NMF defined in Eq. (1) corresponds to simultaneous K-means clustering of the rows (features) and columns (data points) of X, where F can be considered as the clustering indictions for features and G can be considered as the clustering indications for data points. Because co-clustering the both sides of an input data matrix makes use of the interrelatedness between the data points and features, NMF based co-clustering methods usually report superior performance [5, 7]. In the context of Web page classification, the intrinsic linguistic structures of a language is described by X for a set of Web pages, and the prior knowledge by human efforts could be encoded in G, both of which are transformed into F as word (feature) clustering patterns through Eq. (1).

Because two-factor NMF in Eq. (1) is restrictive, which often gives a rather poor low-rank matrix approximation, we may introduce one more factor $S \in \mathbb{R}_+^{k_1 \times k_2}$ to absorb the different scales of X, F and G, which leads to NMTF [7] minimizing the following objective:

$$J_{\text{NMTF}} = \|X - FSG^T\|^2 \quad s.t. \quad F \geq 0, G \geq 0, S \geq 0, \quad (2)$$

where $F \in \mathbb{R}_+^{p \times k_1}$ and $G \in \mathbb{R}_+^{n \times k_2}$. S provides increased degrees of freedom such that the low-rank matrix representation remains accurate while F gives row clusters and G gives column clusters. Most importantly, S is a condensed view of X [13] and represents the associations between word clusters and Web page clusters [25].

Obviously, F and S convey two types of transformed knowledge, which we exactly expect to transfer across languages as outlined in Section 1. However, Eqs. (1–2) are both designed for one single data set, while in cross-language Web page classification we have two separate data sets, one in auxiliary language and the other in target language. Moreover, both of them are unsupervised where prior labeling knowledge is not utilized. Therefore, we further develop NMTF in Eq. (2) and propose a novel joint NMTF approach to transfer knowledge across languages to address the challenges in cross-language Web page classification to achieve improved classification performance.

3. DUAL KNOWLEDGE TRANSFER USING NMTF

In this section, we develop a novel NMTF based DKT approach for cross-language Web page classification, which transfers knowledge from the auxiliary language to the target one by two paths: (1) word cluster approximations and (2) the associations between word clusters and Web page classes. An efficient algorithm to solve the proposed objective with rigorous convergence proof is presented.

3.1 Problem Formalization

For a cross-language Web page classification problem, we have two Web page data sets, one in the auxiliary language $X_a = \left[\mathbf{x}_a^1, \ldots, \mathbf{x}_a^{n_a}\right] \in \mathbb{R}_+^{m \times n_a}$ and the other in the target language $X_t = \left[\mathbf{x}_t^1, \ldots, \mathbf{x}_t^{n_t}\right] \in \mathbb{R}_+^{m \times n_t}$, where $\mathbf{x}_a^i$ represents a Web page in the auxiliary language and $\mathbf{x}_t^i$ represents that in the target language. Thus X_a and X_t can be seen as the document-word co-occurrence matrices of the auxiliary data and target data respectively, or their *tf-idf* normalized counterparts. We assume that the both data sets are using a same vocabulary with m words: if the vocabularies differ, we may simply pad zeros in the feature vectors and re-express them under a same unified vocabulary so that the indices of the feature vectors from the both data sets correspond to the same word. Let $V \in \mathbb{R}^{m \times m}$ be a diagonal matrix with $V_{(ii)} = 1$ if the i-th word occurs in the both data sets, and $V_{(ii)} = 0$ otherwise.

Typically a large amount of Web pages in the auxiliary language are manually labeled, which can be described by an indication matrix $Y_a \in \mathbb{R}^{n_a \times k_2}$ such that $Y_{a(ik)} = 1$ if $\mathbf{x}_a^i$ belongs to the k-th class, and $Y_{a(ik)} = 0$ otherwise. Sometimes, though not always, we also have a limited number of labeled Web pages in target language. We similarly describe them using $Y_t \in \mathbb{R}^{n_t \times k_2}$ such that $Y_{t(ik)} = 1$ if $\mathbf{x}_t^i$ belongs to the k-th class, and $Y_{t(ik)} = 0$ otherwise. Again, we assume that the two data sets share a same set of classes. If not, we can pad the zero columns to Y_a or Y_t, or the both, such that the column indices of the both matrices correspond to the same classes. Our goal is to predict labels for the unlabeled Web pages in the target data set.

Table 1: Some frequently used notations.

X_a	data matrix of Web pages in auxiliary language
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
F_a	word cluster indicator matrix of X_a
F_t	word cluster indicator matrix of X_t
S	the matrix associating word clusters and classes
G_a	class indicator matrix of Web pages in auxiliary language
G_t	class indicator matrix of Web pages in target language
Y_a	true label matrix of Web pages in auxiliary language
Y_t	true label matrix of Web pages in target language
V	word sharing indication matrix of the two data sets
C_a	label indication matrix of X_a
C_t	label indication matrix of X_t

Throughout this paper, we denote the real number set as $\mathbb{R}$ and the nonnegative real number set as $\mathbb{R}_+$. The element at the i-th row and j-th column of a matrix M is denoted as $M_{(ij)}$. Frequently used notations are summarized in Table 1.

3.2 Objective of the Proposed Approach

Given the Web page data X_a in auxiliary language and their labels Y_a, adopting the idea of NMTF, we may factorize X_a by minimizing the following objective [25]:

$$\begin{aligned} J_1 = & \|X_a - F_a S_a G_a^T\|^2 + \alpha \, \mathbf{tr}\left[(G_a - Y_a)^T C_a (G_a - Y_a)\right], \\ & s.t. \quad F_a \geq 0, S_a \geq 0, G_a \geq 0, \end{aligned} \quad (3)$$

where $\mathbf{tr}(\cdot)$ denote the trace of a matrix. In Eq. (3), $\alpha > 0$ is a parameter that determines to which extent we enforce the prior labeling knowledge in the auxiliary language, *i.e.*, $G_a \approx Y_a$. $C_a \in \mathbb{R}^{n_a \times n_a}$ is a diagonal matrix with $C_{a(ii)} = 1$ if $\mathbf{x}_a^i$ is labeled by the i-th row of Y_a, and $C_{a(ii)} = 0$ otherwise. Note that, if $C = I$, all the Web pages in auxiliary language are completely labeled and specified by Y_a.

Transfer knowledge via word cluster approximations by F_a and F_t. Solving the optimization problem in Eq. (3), F_a and S_a contains information of the data in the auxiliary language which is to be transferred to those in the target language. We achieve this by minimizing:

$$\begin{aligned} J_2 = & \|X_t - F_t S_t G_t^T\|^2 + \mathbf{tr}\Big[\beta (G_t - Y_t)^T C_t (G_t - Y_t) \\ & + \gamma (F_t - {F_a}^*)^T V (F_t - {F_a}^*)\Big], \\ & s.t. \quad F_t \geq 0, S_t \geq 0, G_t \geq 0, \end{aligned} \quad (4)$$

where ${F_a}^*$ is obtained by solving Eq. (3). The second term in Eq. (4) acts same as that in Eq. (3), which enforces labeling information in the target domain if it is available. Here, $\beta > 0$, and $C_t \in \mathbb{R}^{n_t \times n_t}$ is a diagonal matrix whose entry $C_{t(ii)} = 1$ if Web page $\mathbf{x}_t^i$ is labeled by the i-th row of Y_t, and $C_{t(ii)} = 0$ otherwise. When the labels for all the Web pages in target language are not available, we have $C_t = 0^{n_t \times n_t}$, which is a zero matrix. The key part is the third term. It enforces the constraint that word clusters in X_t is approximately close to F_a, learned from X_a. The extent of this approximation is determined by $\gamma > 0$. As a result, the label information contained G_a of X_a is transferred to the label assignments G_t in X_t via the semantic word structures F_a and F_t, which is schematically shown red paths in Figure 2.

In order to demonstrate the usefulness of knowledge transfer via word cluster approximations, we give more theoretical analysis on the example in Figure 1. Suppose the knowledge in auxiliary language is certain, we may set γ in Eq. (4) as ∞. In order to see the real effect of prior knowledge to improve classification performance, we temporarily ignore the training information in the target language by minimizing:

$$J_2' = \|X_t - F_a S_t G_t^T\|^2, \tag{5}$$

which is identical to the following problem [5, 7]:

$$\max_{G_t} \ \mathbf{tr}\left(G_t^T X_t^T F_a F_a^T X_t G_t\right) \ . \tag{6}$$

By the equivalence between K-means clustering and principal component analysis (PCA) [4, 24], the clustering of Eq. (6) uses $X_t^T F_a F_a^T X_t$ as the pairwise similarity, whereas K-means clustering uses $X_t^T X_t$ as the pairwise similarity. For the example in Figure 1, we have

$$X_t^T X_t = \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix}, \tag{7}$$

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

Now, with the knowledge F_a learned from auxiliary language, we have

$$X_t^T F_a F_a^T X_t = \begin{bmatrix} 1 & 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 \end{bmatrix}, \tag{8}$$

where we assume we already learned F_a from auxiliary language, which is

$$F_a^T = 2^{-1/2} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{9}$$

Clearly, using the similarity in Eq. (8), K-means clustering will generate (W1, W2) as a cluster and (W3, W4) as another cluster, which is more meaningful as in Figure 1(c).

We may see more directly how knowledge in the word space from auxiliary language is transformed into the Web page space in target language. Let the square root of the semi-definite positive matrix be P: $F_a F_a^T = P^T P$. We have $X_t^T F_a F_a^T X_t = (PX)^T (PX)$ which means we cluster the Web pages using the transformed data $\tilde{X}_t = PX = \left(F_a F_a^T\right)^{1/2} X_t$.

For the example in Figure 1, we have

$$\tilde{X}_t = 2^{-1/2} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} X = 2^{-1/2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}. \tag{10}$$

It is obvious that on this transformed data, W1 and W2 will be clustered into one cluster, W3 and W4 will be clustered into another cluster. This analysis shows how the knowledge in the word space learned from auxiliary language is transformed into the Web page space in target language.

Transfer knowledge via the association between the word clusters and Web page classes by S. As discussed earlier in Section 1, compared to words, the association between word clusters and Web pages classes are more reliable to convey semantic relationships across different languages. Formally, we achieve this by minimizing:

$$J_3 = \|X_t - F_t {S_a}^* G_t^T\|^2 \quad s.t. \quad F_t \geq 0, G_t \geq 0, \tag{11}$$

where ${S_a}^*$ is obtained by solving Eq. (3). As a result, ${S_a}^*$, learned from the auxiliary data set, is used as supervision to classify the target data. Namely, ${S_a}^*$ bridges the source and target languages such that prior labeling knowledge can be transferred from the former to the latter, which is schematically shown green paths in Figure 2.

Our optimization objective. Finally, we may combine the three optimization problems Eqs. (3–11) into a joint optimization objective to minimize:

$$\begin{aligned} & J_{\text{DKT}} = \|X_a - F_a S G_a^T\|^2 + \|X_t - F_t S G_t^T\|^2 \qquad (12) \\ & + \alpha\, \mathbf{tr}\left[(G_a - Y_a)^T C_a (G_a - Y_a)\right] \\ & + \mathbf{tr}\left[\beta (G_t - Y_t)^T C_t (G_t - Y_t) + \gamma (F_t - F_a)^T V (F_t - F_a)\right], \\ & \quad s.t. \quad F_a \geq 0, G_a \geq 0, S \geq 0, F_t \geq 0, G_t \geq 0 \ . \end{aligned}$$

In this formulation S is shared in the two matrix factorizations for both auxiliary and target data, which is used as a bridge to transfer knowledge between them. In addition, through the constraint in the last term, the two data sets are connected by F_a and F_t.

Note that, the last term of Eq. (4) only applies to the common words of X_a and X_t, which are encoded by V. When the auxiliary data set and the target data set do not share any word, *i.e.*, $V = 0^{m \times m}$ is a zero matrix, there will be no knowledge transfer through word cluster approximation 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 Eq. (11), because it is decoupled into two independent subproblems, one for auxiliary data and the other for target data. However, these two cases rarely happen simultaneously. As a result, our model is flexible and can always transfer knowledge in Eq. (12) through the either path, or the both.

Upon solving Eq. (12), the class label for the $\mathbf{x}_t^i$ in target language is determined by

$$\left(\mathbf{x}_t^i\right) = \arg\max_k \ G_{t(ik)} \ . \tag{13}$$

Solving Eq. (12) and assign labels to the unlabeled Web pages in the target language using Eq. (13), our cross-domain Web page classification approach is proposed. Because Eq. (12) transfers knowledge in two different paths, we call it as *Dual Knowledge Transfer (DKT)* approach.

3.3 Optimization Procedures

In the following, we derive the solution to Eq. (12) 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 Eq. (12) as follows,

$$
\begin{aligned}
&J\left(F_a, G_a, S, F_t, G_t\right) = \mathbf{tr}\Big(-2X_a^T F_a S G_a^T \\
&+ G_a S^T F_a^T F_a S G_a^T - 2X_t^T F_t S G_t^T + G_t S^T F_t^T F_t S G_t^T \\
&+ \alpha G_a^T C_a G_a - 2\alpha G_a^T C_a Y_a + \beta G_t^T C_t G_t - 2\beta G_t^T C_t Y_t \\
&+\gamma F_t^T V F_t - 2\gamma F_t^T V F_a + \gamma F_a^T V F_a\Big) ,
\end{aligned}
\tag{14}
$$

where constant terms are discarded.

Computation of F_a. For the constraint $F_a \geq 0$, following standard theory of constrained optimization, we introduce the Lagrangian multiplier $U \in \mathbb{R}^{m \times k_1}$, thus the Lagrangian function is

$$
L\left(F_a\right) = J - \mathbf{tr}\left(U F_a^T\right) \quad . \tag{15}
$$

Setting $\partial L\left(F_a\right)/\partial F_a = 0$, we obtain

$$
U = -2X_a G_a S^T + 2F_a S G_a^T G_a S^T - 2\gamma V F_t + 2\gamma V F_a \quad . \tag{16}
$$

Using Karush-Kuhn-Tucker condition $U_{(ij)} F_{a(ij)} = 0$, we get

$$
\begin{aligned}
&\Big(-2X_a G_a S^T + 2F_a S G_a^T G_a S^T \\
&\quad - 2\gamma V F_t + 2\gamma V F_a\Big)_{(ij)} F_{a(ij)} = 0,
\end{aligned}
\tag{17}
$$

which leads to the following updating formula:

$$
F_{a(ij)} \leftarrow F_{a(ij)} \sqrt{\frac{\left(X_a G_a S^T + \gamma V F_t\right)_{(ij)}}{\left(F_a S G_a^T G_a S^T + \gamma V F_a\right)_{(ij)}}} \quad . \tag{18}
$$

Computation of G_a, S, F_t and G_t. Following the same derivations in Eqs. (15–18), we obtain the updating rules for the rest variables of J_{DKT} as following:

$$
G_{a(ij)} \leftarrow G_{a(ij)} \sqrt{\frac{\left(X_a^T F_a S + \alpha C_a Y_a\right)_{(ij)}}{\left(G_a S^T F_a^T F_a S + \alpha C_a G_a\right)_{(ij)}}} \tag{19}
$$

$$
S_{(ij)} \leftarrow S_{(ij)} \sqrt{\frac{\left(F_a^T X_a G_a + F_t^T X_t G_t\right)_{(ij)}}{\left(F_a^T F_a S G_a^T G_a + F_t^T F_t S G_t^T G_t\right)_{(ij)}}} \tag{20}
$$

$$
F_{t(ij)} \leftarrow F_{t(ij)} \sqrt{\frac{\left(X_t G_t S^T + \gamma V F_a\right)_{(ij)}}{\left(F_t S G_t^T G_t S^T + \gamma V F_t\right)_{(ij)}}} \tag{21}
$$

$$
G_{t(ij)} \leftarrow G_{t(ij)} \sqrt{\frac{\left(X_t^T F_t S + \beta C_t Y_t\right)_{(ij)}}{\left(G_t S^T F_t^T F_t S + \beta C_t G_t\right)_{(ij)}}} \tag{22}
$$

In summary, we present the iterative multiplicative updating algorithm of optimizing Eq. (12) in Algorithm 1.

3.4 Analysis of Algorithm Convergence

In this section, we will investigate the convergence of Algorithm 1. We use the auxiliary function approach [12] to prove the convergence of the algorithm.

Lemma 1. *[12] $Z(h, h')$ is an auxiliary function of $F(h)$ if the conditions $Z(h, h') \geq F(h)$ and $Z(h, h') = F(h)$ are satisfied. [12] If Z is an auxiliary function for F, then F is non-increasing under the update $h^{(t+1)} = \arg\min_h Z(h, h')$.*

Algorithm 1: Algorithm to solve J_{DKT} in Eq. (12).

```
Input: 1. Data matrix $X_a$ in auxiliary language,
2. data matrix $X_t$ in target language,
3. labels of Web pages in auxiliary language $Y_a$,
4. optional labeling information $Y_t$ in target data,
5. trade-off parameters $\alpha$, $\beta$ and $\gamma$.
Initialize $F_a$, $G_a$, $S$, $F_t$ and $G_t$ following [25];
while not converge do
    1. Update $F_a$ using Eq. (18),
    2. Update $G_a$ using Eq. (19),
    3. Update $S$ using Eq. (20),
    4. Update $F_t$ using Eq. (21),
    5. Update $G_t$ using Eq. (22),
end
Predict labels for $\mathbf{x}_t^i$ using Eq. (13).
Output: Labels assigned to the unlabeled Web page $\mathbf{x}_t^i$
        in target language.
```

Lemma 2. *[7] For any matrices $A \in \mathbb{R}_+^{n \times n}$, $B \in \mathbb{R}_+^{k \times k}$, $S \in \mathbb{R}_+^{n \times k}$ and $S' \in \mathbb{R}_+^{n \times k}$, and A and B are symmetric, the following inequality holds*

$$
\sum_{ip} \frac{\left(A S' B\right)_{ip} S_{ip}^2}{S'_{ip}} \geq \boldsymbol{tr}\left(S^T A S B\right) \quad . \tag{23}
$$

Theorem 1. *Write J in Eq. (14) w.r.t. F_a, we have*

$$
\begin{aligned}
J\left(F_a\right) = \boldsymbol{tr}\Big(&-2X_a^T F_a S G_a^T + G_a S^T F_a^T F_a S G_a^T \\
&-2\gamma F_t^T V F_a + \gamma F_a^T V F_a\Big) ,
\end{aligned}
\tag{24}
$$

then the following function

$$
\begin{aligned}
H\left(F_a, F_a'\right) = &- 2\sum_{ij}\left(X_a G_a S^T\right)_{(ij)} F'_{a(ij)} \left(1 + \log\frac{F_{a(ij)}}{F'_{a(ij)}}\right) \\
&+ \sum_{ij}\left(F'_a S G_a^T G_a S^T\right)_{(ij)} \frac{{F_a}^2_{(ij)}}{F'_{a(ij)}} \\
&- 2\gamma \sum_{ij} \left(V F_t\right)_{(ij)} F'_{a(ij)} \left(1 + \log\frac{F_{a(ij)}}{F'_{a(ij)}}\right) \\
&+ \gamma \sum_{ij} \left(V F'_a\right)_{(ij)} \frac{{F_a}^2_{(ij)}}{F'_{a(ij)}}
\end{aligned}
\tag{25}
$$

is an auxiliary function for $J\left(F_a\right)$. Furthermore, it is a convex function in F_a and its global minimum is

$$
F_{a(ij)} = F_{a(ij)} \sqrt{\frac{\left(X_a G_a S^T + \gamma V F_t\right)_{(ij)}}{\left(F_a S G_a^T G_a S^T + \gamma V F_a\right)_{(ij)}}} \quad . \tag{26}
$$

Proof. According to Lemma 2, we have

$$
\mathbf{tr}\left(G_a S^T F_a^T F_a S G_a^T\right) \leq \sum_{ij}\left(F'_a S G_a^T G_a S^T\right)_{(ij)} \frac{{F_a}^2_{(ij)}}{F'_{a(ij)}}, \tag{27}
$$

$$
\mathbf{tr}\left(F_a^T V F_a\right) \leq \sum_{ij}\left(V F'_a\right)_{(ij)} \frac{{F_a}^2_{(ij)}}{F'_{a(ij)}} \quad . \tag{28}
$$

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

$$
\mathbf{tr}\left(X_a^T F_a S G_a^T\right) \geq \sum_{ij}\left(X_a G_a S^T\right)_{(ij)} F'_{a(ij)} \left(1 + \log\frac{F_{a(ij)}}{F'_{a(ij)}}\right), \tag{29}
$$

$$\mathbf{tr}\left(F_t^T V F_a\right) \geq \sum_{ij} \left(VF_t\right)_{(ij)} F'_{a(ij)} \left(1+\log\frac{F_{a(ij)}}{F'_{a(ij)}}\right) . \tag{30}$$

Summing over all the bounds in Eqs. (27–30), we can obtain $H\left(F_a, F'_a\right)$, which clearly satisfies (1) $H\left(F_a, F'_a\right) \geq J\left(F_a\right)$ and (2) $H\left(F_a, F_a\right) = J\left(F_a\right)$.

Then, fixing F'_a, we minimize $H\left(F_a, F'_a\right)$.

$$\frac{\partial H\left(F_a, F'_a\right)}{\partial F_{a(ij)}} = -2\left[\left(X_a G_a S^T\right)_{(ij)} + \gamma\left(VF_t\right)_{(ij)}\right]\frac{F'_{a(ij)}}{F_{a(ij)}} + 2\left[\left(F'_a S G_a^T G_a S^T\right)_{(ij)} + \gamma\left(VF'_a\right)_{(ij)}\right]\frac{F_{a(ij)}}{F'_{a(ij)}} \tag{31}$$

and the Hessian matrix of $H\left(F_a, F'_a\right)$ is

$$\frac{\partial^2 H\left(F_a, F'_a\right)}{F_{a(ij)}F_{a(kl)}} = \delta_{ik}\delta_{jl}\left\{2\left[\left(X_a G_a S^T\right)_{(ij)} + \gamma\left(VF_t\right)_{(ij)}\right]\frac{F'_{a(ij)}}{F_{a(ij)}^2} +2\left[\left(F'_a S G_a^T G_a S^T\right)_{(ij)} + \gamma\left(VF'_a\right)_{(ij)}\right]F'_{a(ij)}\right\}, \tag{32}$$

which is a diagonal matrix with positive diagonal elements. Therefore $H\left(F_a, F'_a\right)$ is a convex function of F_a, and we can obtain the global minimum of $H\left(F_a, F'_a\right)$ by setting $\partial H\left(F_a, F'_a\right)/\partial F_{a(ij)} = 0$ and solving for F_a, from which we get Eq. (26). This completes the proof of Theorem 1. □

Theorem 2. *Using Algorithm 1 to update F_a, $J\left(F_a\right)$ in Eq.* (24) *will monotonically decreases.*

Proof. By Lemma 1 and Theorem 1, we can get that $J\left({F_a}^0\right) = H\left({F_a}^0, {F_a}^0\right) \geq H\left({F_a}^1, {F_a}^0\right) \geq J\left({F_a}^1\right) \ldots$ Therefore $J\left(F_a\right)$ is monotonically decreasing. □

Theorem 3. *Using Algorithm 1 to update G_a, S, F_t and G_t, the respective objectives will monotonically decrease.*

Theorem 3 can be similarly proved as Theorems (1–2).

Because J in Eq. (12) is obviously lower bounded by 0, Algorithm 1 is guaranteed to converge by Theorems (2–3).

4. RELATED WORKS

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

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 where training and test data are obtained from different resources thereby in different distributions [13–16, 18–21, 23, 25]. For a comprehensive survey of transfer learning, we refer readers to [17].

Cross-language classification. Cross-language Web page and document classification has attracted increased attention in recent years due to its importance in information retrieval. Bel *et al.* [1] 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 first strategy. [16] 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.* [15] classify Chinese Web pages using English data source by utilizing the information bottleneck theory. Other cross-language text classification researches include [23] (Chinese-English), [19] (English-Spanish-French), [20] (English-Chinese-French), *etc.*, to be mentioned.

Table 2: Description of testing data sets. English is used as auxiliary language in all the testing data sets.

Data	Target language	# Labeled auxiliary data	# Labeled target data
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

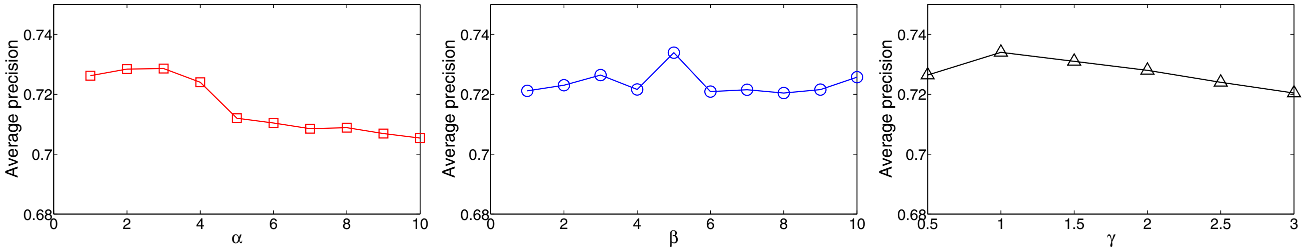
NMTF. NMF is a useful learning method to approximate a nonnegative input data matrix by the product of factor matrices [11,12], which has been applied to solve many real world problems including dimensionality reduction, pattern recognition, clustering and classification [3,5–8,13,14,22,25]. Recently, Ding *et al.* extended NMF [7] to NMTF and explored its relationships to K-means/spectral clustering [5,7]. 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 [3,8,13,14,22,25], among which [13] and [25] 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 word cluster approximations. 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 [13]. 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 [13] and [25], but enjoys the advantages of both of them, with additional flexibilities to allow training data appearing in various forms.

5. EXPERIMENTS

In this section, we evaluate the proposed Dual Knowledge Transfer (DKT) approach in cross-language Web page classification, and compare it with several state-of-the-art supervised, semi-supervised and transfer learning classifiers.

5.1 Data Preparation

We conduct our empirical evaluations on a publicly avail-



(a) α. (b) β. (c) γ.

Figure 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.

able multi-lingual Web page data set — cross lingual sentiment corpus[2] [18]. 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 4 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.* [2]. All the Web pages in the corpus are divided into three categories upon the product they describe: books, DVDs and music.

In our experiments, we randomly pick up 5,000 Web pages from each language. Same as [18], we use English as auxiliary language and the rest three as target languages separately. Therefore we end up with three language pairs for testing: English-Germen, English-French, and English-Japanese. Because in real world applications not all the Web pages in auxiliary language are labeled, we randomly pick up 70% of English Web pages for each class as labeled data. On the other hand, because in real world applications the Web pages in target language are mostly unlabeled, we simulate two different cases: (1) no labeled Web pages in 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 2. For each testing data set, our task is to classify the unlabeled Web pages in the corresponding target language.

5.2 Performance Comparisons

We compare the proposed DKT approach against the supervised learning method (1) Support Vector Machine (SVM) method, and the semi-supervised learning method (2) Transductive SVM (TSVM) method [9] as baselines. We also compare to the two closely related cross-domain learning methods based on NMTF: (3) Knowledge Transformation by Words (KTW) method [13], (4) Matrix Tri-factorization based classification framework (MTrick) [25]; and a very recent cross-language Web classification method using information bottleneck theory (IB) [15].

5.2.1 Experimental Setups

SVM and TSVM methods can use either the labeled data in target language or the labeled data in both auxiliary and target language. We therefore 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 laguanges 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 ($i.e.$, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$) where γ is set as $1/m$. SVM and TSVM are implemented by SVM^{light} [10].

The parameters of KTW and MTrick are set as optimal following their original works [13, 25]. The iteration number of IB method is set as 100.

For our approach, due to the nature of our optimization objective in Eq. (12), we always use S, $i.e.$, 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)", $i.e.$, set $\gamma = 0$ in Eq. (12); and (2) use words transfer denoted as "DKT". Upon some preliminary test, for our appraoch we set the tradeoff parameters $\alpha = \beta = 1$, and $\gamma = 1.5$, the number of word clusters is set to same as Web page classes $k_1 = k_2 = 3$, the error threshold in Algorithm 1 to determine convergence is set $\varepsilon = 10^{-11}$, and the maximum iterating number is 100.

5.2.2 Evaluation Metrics

Two widely used classification performance metrics in statistical learning and information retrieval 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:

$$P = \frac{1}{\mathcal{C}} \sum_{c \in \mathcal{C}} \frac{\{d | d \in X_c \wedge h(d) = f(d) = c\}}{\{d | d \in X_c \wedge h(d) = c\}} \tag{33}$$

$$R = \frac{1}{\mathcal{C}} \sum_{c \in \mathcal{C}} \frac{\{d | d \in X_c \wedge h(d) = f(d) = c\}}{\{d | d \in X_c \wedge f(d) = c\}} \tag{34}$$

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

$$F_1 = \frac{2PR}{P + R} \tag{35}$$

[2] `http://www.webis.de/research/corpora/`

5.2.3 *Experimental Results*

Table 3 presents the classification performances measured by macro-average precision and F_1 score of the compared methods on six different test data sets. From the results we have the following observations. First, the proposed DKT approach consistently outperforms the other compared methods. DKT (S only) method is always worse than DKT approach, which confirms the usefulness of the knowledge transfer path by word cluster approximations. Second, 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. Third, our DKT approach is able to transfer knowledge in two paths, *i.e.*, word cluster approximations and the associations between word clusters and Web page classes, thus it achieves encouraging classification performance on all the six test data sets. In contrast, KTW method can only transfer knowledge through word cluster approximations, 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 least, our approach is able to exploit labeling information in both auxiliary and target data, whereas KTW method and MTrick method cannot benefit from labeling information in target domain, and SVM_T method and TSVM_T method cannot work with labeling information in auxiliary data. The more labeled data in target domain, the better classification performance our approach can achieve. In summary, all the above observations demonstrate the effectiveness of the proposed DKT approach in cross-language Web page classification.

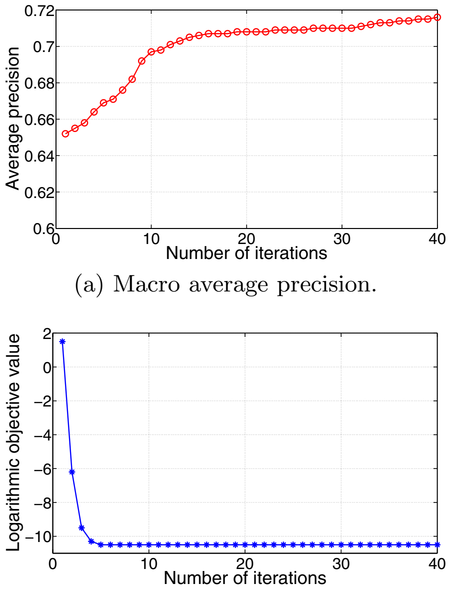
5.3 Parameter Effect

The proposed DKT approach has three parameters α, β, γ in Eq. (12). Although it is tedious to seek an optimal combination of them, we can demonstrate that the performance of our DKT approach is not sensitive when the parameters are sampled in some value ranges. We bound the parameters in the ranges of $1 \leq \alpha \leq 10$, $1 \leq \beta \leq 10$ and $0.5 \leq \gamma \leq 3$ upon preliminary tests and evaluate them on data set D2 as it has labeled Web pages in both auxiliary language and target language.

From Figure 3 we can see that the average performance of all the parameter settings is almost the same as the results from the default parameters. Morever, the variance of all the parameter settings is small. It shows that the performance of our approach is stable with respect to the parameters in a considerably large range.

5.4 Algorithm Convergence

Because our DKT approach employs an iterative algorithm, an important issue is its convergence property. In Section 3.4, we have already theoretically proved the convergence of the algorithm. Here, we empirically check the convergence property of the proposed iterative algorithm. The classification accuracy in each iteration to classify D1 test data set is shown in Figure 4(a), where the x-axis represents the number of iterations, and the y-axis denote the macro-average precision. From the figure, it be can seen that DKT approach converges within about 10 iterations, which indicates that our algorithm is fast. In addition, the objective values of our algorithm in each iteration are plotted in Figure 4(b), which shows that the objective value of our algorithm keep to decrease along with iterative process, which is in accordance with our theoretical analysis.



(a) Macro average precision.

(b) Objective values.

Figure 4: Number of iterations *vs.* the performance of the proposed approach measured by macro-average precision and Objective Value.

6. CONCLUSIONS

In this paper, we proposed a novel NMTF based DKT approach for cross-language Web page classification. Our approach adopts the idea of transfer learning to pass knowledge across languages, instead of simply combining the Web page data from different languages. By carefully examine the cross-language Web page classification problem, we observed that common semantic patterns usually exist in Web pages for a same topic from different languages. Moreover, we also observed 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 paths: word cluster approximations 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. Extensive experiments using a real world cross-language Web page data set demonstrated encouraging results from a number of aspects that validate our approach.

Acknowledgments

This research was supported by NSF-CNS 0923494, NSF-IIS 1041637, NSF-CNS 1035913.

Table 3: Macro-average precision and F_1 measure for each classifiers on each test data sets. The results demonstrate the advantage of the proposed DKT approach.

Data	Metrics	Compared methods								
		SVM_T	SVM_TS	TSVM_T	TSVM_TS	KTW	MTrick	IB	DKT (S only)	DKT
D1	Precision	–	0.682	–	0.689	0.673	0.695	0.691	0.697	0.716
	F_1	–	0.479	–	0.483	0.481	0.490	0.492	0.495	0.508
D2	Precision	0.679	0.692	0.682	0.701	0.675	0.699	0.703	0.718	0.730
	F_1	0.468	0.481	0.475	0.489	0.483	0.492	0.501	0.505	0.510
D3	Precision	–	0.670	–	0.675	0.663	0.682	0.680	0.683	0.701
	F_1	–	0.470	–	0.475	0.470	0.481	0.480	0.483	0.498
D4	Precision	0.663	0.682	0.670	0.687	0.669	0.681	0.690	0.702	0.718
	F_1	0.452	0.469	0.461	0.472	0.471	0.481	0.486	0.492	0.501
D5	Precision	–	0.662	–	0.668	0.656	0.674	0.672	0.679	0.688
	F_1	–	0.463	–	0.468	0.462	0.472	0.470	0.477	0.485
D6	Precision	0.651	0.676	0.663	0.676	0.658	0.672	0.681	0.695	0.707
	F_1	0.447	0.460	0.456	0.467	0.463	0.475	0.478	0.486	0.493

7. REFERENCES

[1] N. Bel, C. Koster, and M. Villegas. Cross-lingual text categorization. *Research and Advanced Technology for Digital Libraries*, pages 126–139, 2004.

[2] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.

[3] G. Chen, F. Wang, and C. Zhang. Collaborative filtering using orthogonal nonnegative matrix tri-factorization. *Information Processing and Management*, 45(3):368–379, 2009.

[4] C. Ding and X. He. K-means clustering via principal component analysis. In *ICML*, 2004.

[5] C. Ding, X. He, and H. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, 2005.

[6] C. Ding, T. Li, and M. Jordan. Convex and semi-nonnegative matrix factorizations. *TPAMI*, 32(1):45–55, 2010.

[7] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *SIGKDD*, 2006.

[8] Q. Gu and J. Zhou. Co-clustering on manifolds. In *SIGKDD*, 2009.

[9] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209.

[10] T. Joachims. SVMLight: Support Vector Machine. *http://svmlight.joachims.org/*, 1999.

[11] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[12] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

[13] T. Li, V. Sindhwani, C. Ding, and Y. Zhang. Knowledge transformation for cross-domain sentiment classification. In *SIGIR*, 2009.

[14] T. Li, V. Sindhwani, C. Ding, and Y. Zhang. Bridging Domains with Words: Opinion Analysis with Matrix Tri-factorizations. In *SDM*, 2010.

[15] X. Ling, G. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu. Can Chinese web pages be classified with English data source? In *WWW*, 2008.

[16] J. Olsson, D. Oard, and J. Hajič. Cross-language text classification. In *SIGIR*, 2005.

[17] S. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 2009.

[18] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *ACL*, 2010.

[19] G. Ramírez-de-la Rosa, M. Montes-y Gómez, L. Villaseñor Pineda, D. Pinto-Avendaño, and T. Solorio. Using information from the target language to improve crosslingual text classification. In *Proceedings of the 7th international conference on Advances in natural language processing*, 2010.

[20] L. Shi, R. Mihalcea, and M. Tian. Cross language text classification by model translation and semi-supervised learning. In *EMNLP*, 2010.

[21] X. Wan. Co-training for cross-lingual sentiment classification. In *ACL*, 2009.

[22] F. Wang, T. Li, and C. Zhang. Semi-supervised clustering via matrix factorization. In *SDM*, 2008.

[23] K. Wu and B. Lu. A refinement framework for cross language text categorization. In *AIRS*, 2008.

[24] H. Zha, X. He, C. Ding, H. Simon, and M. Gu. Spectral relaxation for k-means clustering. In *NIPS*, 2001.

[25] F. Zhuang, P. Luo, H. Xiong, Q. He, Y. Xiong, and Z. Shi. Exploiting associations between word clusters and document classes for cross-domain text categorization. In *SDM*, 2010.