Correlation-Based Cross-Modal Information Retrieval Research Postprint

Authors: Ding Heng, Lu Wei

Date: 2017-10-11T00:00:00+00:00

Abstract

【Objective】 Review the fundamental strategies and core issues in correlation-based cross-modal information retrieval, and explore the advantages and disadvantages of using Partial Least Squares for feature subspace projection from the perspective of improving retrieval performance.

【Method】 On the Wikipedia cross-modal information retrieval dataset, LDA and BOW models are respectively adopted as feature representation methods for text and image resources, cosine distance is used as the similarity measurement method, and least squares method is employed to learn the feature subspace projection function instead of Canonical Correlation Analysis.

【Results】 Through comparative analysis of the impact of three feature subspace projection methods—Canonical Correlation Analysis, Partial Least Squares Regression, and Partial Least Squares Correlation—on cross-modal information retrieval results using three retrieval evaluation metrics: P@K, MAP, and NDCG, the results indicate that Partial Least Squares Correlation achieves the best performance.

【Limitations】 Partial Least Squares assumes linear relationships between data and orthogonal relationships between data basis vectors when processing data, thus it cannot address nonlinear and non-orthogonal problems.

【Conclusion】 The feature subspace projection learned using Partial Least Squares Correlation demonstrates stronger consistency with the original space information, yielding more stable cross-modal information retrieval results.

Full Text

A Study on Correlation-based Cross-Modal Information Retrieval

Ding Heng[1], Lu Wei[1],[2]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China)

Abstract

[**Objective**] This study systematically reviews the fundamental strategies and core issues in correlation-based cross-modal information retrieval, and examines the advantages and disadvantages of using partial least squares for feature subspace projection to improve retrieval effectiveness. [**Methods**] Using the Wikipedia cross-modal information retrieval dataset, we employed LDA and BOW models as feature representations for text and image resources respectively, with cosine distance as the similarity metric, and utilized least squares methods to learn feature subspace projection functions as an alternative to canonical correlation analysis. [**Results**] Through comparative analysis of three feature subspace projection methods—canonical correlation analysis, partial least squares regression, and partial least squares correlation—on cross-modal retrieval results using three evaluation metrics (P@K, MAP, and NDCG), the results demonstrate that partial least squares correlation achieves the best performance. [**Limitations**] Partial least squares assumes linear relationships between data and orthogonality between basis vectors, thus cannot address non-linear and non-orthogonal problems. [**Conclusions**] Feature subspace projection learned using partial least squares correlation demonstrates stronger consistency with original spatial information, yielding more stable cross-modal information retrieval results.

Keywords: Cross-Modal Information Retrieval; Partial Least Squares; Subspace Projection

---

Introduction

With advances in multimedia technology and increasing diversification of information resources, traditional information retrieval techniques have undergone significant transformation, evolving from text-based approaches toward content-based multimedia information retrieval. Research on content-based image retrieval [**?**], fingerprint-based music retrieval [**?**], and content-based video retrieval [**?**] has matured, leading to commercial applications such as "search by image" and "query by humming" that effectively address information retrieval within homogeneous modality spaces. However, retrieval systems often face

requirements like "a user has a bird photo and wants to find related textual descriptions, video, and audio clips," which can be reduced to the problem of enabling mutual retrieval of information resources across different modality spaces (text, video, audio, image, etc.).

Current information retrieval systems primarily rely on content-based multimedia retrieval techniques to find relevant resources within the same modality space, then integrate target modality information from these resources to return result lists. For instance, in "web search by image," the system first performs "image-to-image" search to find web pages containing similar images, then returns text information associated with those images. This approach suffers from two major limitations: it cannot retrieve relevant text from pages containing no images, and text associated with similar images may not actually be relevant to the query image. Cross-modal information retrieval attempts to directly establish associations between information resources across different modality spaces to overcome these deficiencies.

Cross-modal information retrieval (also called cross-media information retrieval) represents a relatively new research area in multimedia information retrieval, involving multimedia information representation, heterogeneous feature association mining, subspace projection, semantic inference, and related technologies. It achieves transformation of information representation across multiple modality spaces by establishing mappings between modalities, ultimately supporting retrieval that transcends modality differences (heterogeneous data types). Through designed cross-modal retrieval experiments and using three standard information retrieval evaluation metrics as benchmarks, this study explores the advantages and disadvantages of different multivariate statistical analysis methods for processing heterogeneous feature information and performing feature subspace projection. The main contributions and innovations of this paper are twofold: first, systematically reviewing and summarizing the core steps and strategies in correlation-based cross-modal information research; second, proposing the use of partial least squares to mine heterogeneous feature associations for the subspace projection step, with experimental results confirming that partial least squares is more suitable than traditional canonical correlation analysis for correlation-based cross-modal information retrieval frameworks.

2.1 Multimedia Information Processing

Multimedia information processing technologies have been widely applied across many research domains. For example, reference [**?**] clustered local invariant image features into visual words and combined image region semantic information with bag-of-words models using spatial pyramid models to achieve image scene semantic analysis and understanding. Reference [**?**] applied Latent Dirichlet Allocation (LDA) to model short texts, addressing feature sparsity and contextual dependency while exploring short text semantic understanding at the topic level. Reference [**?**] applied Hierarchical Dirichlet Process (HDP) to search engine user log analysis, clustering verbs and dependent nouns in queries to understand

user search intent semantics. Reference [**?**] used Fast Combinatorial Hashing algorithms for music information modeling based on signal spectrum analysis, enabling audio information retrieval via "music fingerprints." These multimedia information processing technologies enable applications such as homogeneous information retrieval and recommendation, and can represent semantic content of information resources to some extent.

2.2 Cross-Media Semantic Information Mining

However, multimedia information processing technologies fail to bridge heterogeneous features of information resources. Consequently, researchers have exploratorily investigated intrinsic connections between cross-media information. Reference [**?**] noted that features of the same information resource under different modalities possess certain latent connections, and used canonical correlation analysis to model associations between heterogeneous data (audio and image), transforming different modality resources into a common subspace to enable cross-media information measurement. Reference [**?**] proposed using singular value decomposition and latent semantic indexing for cross-media semantic relationship modeling, and compared the effectiveness of singular value decomposition, latent semantic indexing, and canonical correlation analysis in mining heterogeneous feature relationships through cross-media retrieval experiments. References [**?**, **?**] incorporated ontology technology into cross-media information processing, constructing semantic associations between multimedia information through relationship-based knowledge reasoning and ontology learning to measure cross-media information differences. Reference [**?**] addressed consistency issues in image and audio content representation, proposing a semi-supervised correlation-preserving mapping algorithm (SSCPM) to mine latent commonalities between image and audio data features. References [**?**, **?**] analyzed differences between low-level and high-level semantic features for cross-modal information retrieval through evaluation results, noting that multi-level feature fusion better represents information commonalities across cross-media data. References [**?**, **?**] proposed spatiotemporal context semantic machine models and proximity graph models, discussing cross-media semantic information mining methods from the perspective of cross-modality correlation propagation, and explored mutual retrieval between text and image information.

The primary approach in these cross-media semantic information mining studies is "constructing an isomorphic semantic subspace to project feature data of different dimensions and scales, thereby enabling relationship measurement of cross-modal information to serve cross-modal information retrieval research." The core problem is learning a subspace that preserves individual modality characteristics while fusing cross-modal information commonalities. Current isomorphic feature subspace construction methods fall into two categories:

1. **Correlation-based feature subspace projection**: This method employs a maximum correlation strategy, primarily using canonical correlation analysis to mine latent correlations between low-level features of

different modality information and learn optimal subspace projection matrices for heterogeneous feature space transformation.

2. **High-level semantic-based feature subspace learning**: This method utilizes machine learning techniques to directly construct isomorphic semantic feature spaces for heterogeneous data at the semantic level through classification algorithms, enabling similarity measurement based on this space.

The second approach heavily relies on multi-class classification algorithm effectiveness. However, classification performance typically decreases as the number of categories increases, limiting the dimensionality of constructible semantic feature spaces and essentially reducing discriminability between retrieval objects. Moreover, expanding semantic feature space dimensions requires relearning classification models and parameter tuning, representing a parameter-dependent solution unsuitable for practical retrieval applications. Therefore, this paper focuses exclusively on optimizing correlation-based cross-modal information retrieval.

3 Correlation-Based Cross-Modal Information Retrieval

We propose that correlation-based cross-modal information retrieval system frameworks consist of three main components: multi-modal information representation, feature subspace projection, and similarity measurement and ranking.

Multi-modal information representation primarily investigates how to encode information resources within the same modality to effectively distinguish individual differences within classes. Formally, multi-modal information representation can be viewed as using mathematical vectors to characterize information resources from different perspectives, where different perspectives manifest as the same information resource being representable by vectors of different dimensions and values. The feature representation of information resources in a specific modality can be formally defined as: for a given information resource set $S = \{S_1, S_2, \ldots, S_n\}$, find an m-dimensional vector space L where each information resource S_i can be represented by some vector in this space. This paper uses LDA topic space and BOW visual word bag space as text and visual feature representations for information resources, respectively.

3.2 Correlation-Based Feature Subspace Projection

Feature subspace projection analyzes latent connections between heterogeneous features of information resources across different modality spaces, thereby projecting heterogeneous data into a common feature subspace to address feature heterogeneity. Correlation-based feature subspace projection mines latent correlations between low-level features of different modality information to learn optimal subspace projection matrices for heterogeneous feature space transformation. The core objective is projecting information resources of different

modalities from heterogeneous feature spaces into an isomorphic feature space to enable direct relationship measurement.

This process can be formally described as: for a given information resource set $S = \{S_1, S_2, \dots, S_n\}$, where S_i has vector representation $\{L_1, L_2, \dots, L_m\}$ in m-dimensional feature space L and vector representation $\{G_1, G_2, \dots, G_n\}$ in n-dimensional feature space G, learn spatial projection relationships φ_L, φ_G and t-dimensional feature subspace O through some strategy F (subspace correlation maximization) or algorithm, such that $(O_1, O_2, \dots, O_t) = \varphi_L(L_1, L_2, \dots, L_m) = \varphi_G(G_1, G_2, \dots, G_n)$. Here φ_L and φ_G are spatial projection functions, and feature subspace O is called the maximum correlation subspace. Its geometric meaning is illustrated in Figure 1 [Figure 1: see original paper].

3.3 Retrieval Ranking Algorithm Based on Feature Subspace

Correlation-based cross-modal information retrieval essentially measures correlation between query information resources and retrieved information resources in isomorphic feature subspace O using some distance calculation method, and ranks them by correlation magnitude. The algorithm pseudocode is as follows:

```
for Sj in S do
    Score(St,Sj)=Dis(Sto, Sjo)
end for
Sort S on Score(St, Sj)
```

Where S_t is any query with vector representation $\{L_1, L_2, \dots, L_m\}$ in feature space L; S is the resource collection to be retrieved, $S_j \in S$, with S_j expressed in feature space G; Dis is the distance calculation formula; and $Score(S_t, S_j)$ represents the relevance score between query S_t and record S_j. Other strategies for similarity measurement and ranking can directly employ distance calculation methods from machine learning, with specific details available in reference [**?**].

3.4 Application Analysis of Partial Least Squares

This paper argues that differences in correlation-based cross-modal information retrieval primarily stem from variations in these three core steps—different strategies within the same step constitute the main cause of retrieval effectiveness differences. Therefore, improvements to any step will enhance cross-modal information retrieval performance. Feature subspace projection is the most critical step in correlation-based cross-modal information retrieval research and currently represents the only approach for fusing feature data of different scales and dimensions. Existing studies [**?**, **?**, **?**] predominantly use canonical correlation analysis to find maximum correlation subspaces for the same information across different modalities as the execution strategy and mathematical solution for this step. However, as a multivariate statistical analysis method, canonical correlation analysis has certain defects in representing relationships between subprojections using linear regression.

Partial least squares, as a second-generation multivariate regression analysis method, simultaneously incorporates advantages of multiple linear regression, principal component analysis, and canonical correlation analysis, and has been widely applied in economics, mechanical control technology, social survey research, chemometrics, neuroimaging, and other fields. Theoretically, partial least squares not only achieves the functionality of canonical correlation analysis but also offers additional advantages such as noise reduction and highlighting major latent variables. Therefore, this study proposes that introducing partial least squares into cross-modal information retrieval frameworks will optimize correlation-based cross-modal information retrieval results. Partial least squares mainly includes partial least squares regression (PLSR) and partial least squares correlation (PLSC), with the former primarily used for prediction and the latter commonly employed for latent variable association mining. Specific mathematical theories and derivations are available in reference [**?**].

The essence of cross-modal information retrieval based on partial least squares is using partial least squares (corresponding to strategy F in Section 3.2 and the feature subspace projection step in Section 4.1) to solve mapping functions φ_L, φ_G from original feature spaces L, G to feature subspace O, highlighting principal component effects and suppressing data noise while maintaining maximum correlation between original features.

To investigate the application of partial least squares in cross-modal information retrieval frameworks, this study designed relevant experiments.

4.1 Experimental Data and Related Processes

Given mature applications of semantic technology in text processing and image analysis, this experiment selected text and image as original information for cross-modal information retrieval, measuring final results through "text-to-image" and "image-to-text" retrieval tasks. The experiment used the Wikipedia cross-modal information retrieval dataset [**?**], which contains 2,866 Wikipedia documents across 10 topics, with each document consisting of a "text-image" pair belonging to a specific topic. Among these, 2,173 documents constitute the training set TRAIN for learning spatial projection functions φ_L, φ_G and feature subspace O, while the remaining 693 documents form the test set TEST for evaluating cross-modal information retrieval ranking algorithm results. Data distribution is shown in Table 1 .

Based on the correlation-based cross-modal information retrieval framework introduced in Section 3, the three core components of this experiment—multi-modal information representation, feature subspace projection, and similarity measurement and ranking—are described as follows (the selection of dimensions for feature spaces L, G, and O has minimal impact on experimental results [**?**]):

1. **Multi-modal information representation**: For document text information, this experiment used the gensim toolkit to extract features in LDA topic space, constructing feature space L with dimension $m = 10$.

For image information, the VLFeat computer vision library calculated features in BOW image semantic space, constructing feature space G with dimension $n = 128$.

2. **Feature subspace projection**: Three experimental groups were established using CCA, PLSR, and PLSC algorithms from the scikit-learn toolkit to learn spatial projection functions φ_L, φ_G and vector representations $S_i = (O_1, O_2, \ldots, O_t)$ for documents S_i in feature subspace O on the training set data, with subspace dimension $t = 9$.

3. **Similarity measurement and ranking**: Using vector cosine similarity as the correlation metric, the relevance score between documents S_t and S_j is calculated as:

$$Score(S_t, S_j) = \frac{S_{to} \cdot S_{jo}}{||S_{to}|| \cdot ||S_{jo}||}$$

4.2 Experimental Results and Analysis

The experiment ultimately performed cross-modal information retrieval on the test set, including "text-to-image"and "image-to-text"tasks, with relevance judgments based on topical consistency between queries and retrieved documents. Retrieval effectiveness was evaluated using three metrics—P@K (Precision at K), MAP (Mean Average Precision), and NDCG (Normalized Discounted Cumulative Gain)—to examine method impacts from multiple perspectives and demonstrate generalizability of retrieval result optimization.

Comparative analysis of P@K values (K=5,10,15,20,30) for "text-to-image" and "image-to-text" tasks using CCA (canonical correlation analysis), PLSR (partial least squares regression), and PLSC (partial least squares correlation) as feature subspace learning algorithms is shown in Figure 2 [Figure 2: see original paper]. The results show PLSC achieves optimal performance in both tasks. In the "text-to-image" task, P@K decreases as K increases, with CCA showing steeper curve slopes while partial least squares-based methods (PLSR, PLSC) exhibit gentler slopes, indicating more stable feature subspace projection learning compared to CCA. In the "image-to-text" task, all three curves show gentle slopes, differing markedly from the "text-to-image"task performance, suggesting that text information projections in feature subspace are discrete and uniformly distributed, while image information projections exhibit clear topic-based clustering characteristics, as illustrated in Figure 3 [Figure 3: see original paper].

MAP scores for the three retrieval experiments are presented in Table 2 , showing PLSC achieves the best results in both tasks. Compared with CCA, performance improved by 19.1% in the "text-to-image" task and 36.7% in the "image-to-text" task, with an average improvement of 28.2%. Two-tailed paired t-tests confirm statistically significant improvements (p_1=0.012, p_2=0.061).

Comparative NDCG score analysis across the three experimental groups for

both tasks is shown in Figure 4 [Figure 4: see original paper]. Examining NDCG scores by topic reveals that the three methods show varying effectiveness across different topics and tasks. However, in terms of overall NDCG scores, PLSC achieves optimal performance in both tasks (NDCG values of 0.2378 and 0.1982 respectively, with average NDCG of 0.2179), representing a 70.7% improvement over CCA. Two-tailed paired t-tests confirm statistically significant improvements (p_1=0.024, p_2=0.036). PLSR shows similar effectiveness to CCA in the "text-to-image" task but outperforms CCA in the "image-to-text" task.

Considering P@K, MAP, and NDCG evaluation metrics comprehensively, partial least squares correlation outperforms canonical correlation analysis across all three metrics, while partial least squares regression demonstrates unstable performance. We therefore conclude that partial least squares correlation is more suitable for correlation-based cross-modal information retrieval theoretical frameworks. Compared with canonical correlation analysis, feature subspace projection learned using partial least squares correlation shows stronger consistency with original spatial information, yielding more stable cross-modal information retrieval results.

Conclusion

Research on content-based multimedia information retrieval has matured, with applications like "search by image" and "query by humming" solving information retrieval problems within homogeneous modality spaces but failing to overcome heterogeneous data type limitations. Cross-modal information retrieval research provides a new solution approach; however, current correlation-based cross-modal information retrieval studies predominantly use canonical correlation analysis to construct feature subspaces, which has certain limitations. This paper introduces partial least squares into the correlation-based cross-modal information retrieval framework and designs corresponding retrieval experiments. Experimental results demonstrate that the partial least squares correlation algorithm effectively optimizes retrieval results.

This study selected text and image data to investigate partial least squares optimization of cross-media correlations between these two different modality information resources. The method is equally applicable to other modality information resources (such as audio, image, video) and cross-language information retrieval research. The primary limitation lies in partial least squares' assumption of linear relationships between data and orthogonality between basis vectors when processing data, thus being unable to solve non-linear and non-orthogonal problems. Future research will focus on non-linear feature subspace learning to compensate for deficiencies caused by partial least squares' linear and orthogonal assumptions.

References

[1] Smeulders A W M, Worring M, Santini S, et al. Content-based Image Re-

trieval at the End of the Early Years [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1349-1380.

[2] Wang A. An Industrial Strength Audio Search Algorithm [C]. In: Proceedings of International Society for Music Information Retrieval Conference, Baltimore, Maryland, USA. 2003: 7-13.

[3] Snoek C G M, Worring M. Concept-based Video Retrieval [J]. Foundations and Trends in Information Retrieval, 2008, 2(4): 215-322.

[4] Wang Yuxin, Guo He, He Changqin, et al. Bag of Spatial Visual Words Model for Scene Classification [J]. Computer Science, 2011, 38(8): 265-268.

[5] Zhang Zhifei, Miao Duoqian, Gao Can. Short Text Classification Using Latent Dirichlet Allocation [J]. Journal of Computer Applications, 2013, 33(6): 1587-1590.

[6] Duan Ruixue, Wang Xiaojie, Sun Yueping, et al. Clustering User Goals Based on Hierarchical Dirichlet Process Topic Model [J]. Journal of Beijing University of Posts and Telecommunications, 2011, 34(S1): 55-58.

[7] Wu F, Zhang H, Zhuang Y. Learning Semantic Correlations for Cross-Media Retrieval [C]. In: Proceedings of IEEE International Conference on Image Processing, Atlanta, USA. IEEE, 2006: 1465-1468.

[8] Zhang Hong, Wu Fei, Zhuang Yueting. Cross-Media Retrieval Method Based on Feature Subspace Learning [J]. Pattern Recognition and Artificial Intelligence, 2008, 21(6): 739-745.

[9] Hu Tao, Wu Gangshan, Ren Tongwei, et al. Ontology-based Cross-media Retrieval Technique [J]. Computer Engineering, 2009, 35(8): 266-268.

[10] Ming Junren, He Chao. Research on Cross-media Retrieval Method in Digital Library Based on Semantic Association Mining [J]. Library and Information Service, 2013, 57(7): 101-105.

[11] Zhang Hong. Correlation Mining Based Cross-media Retrieval [D]. Hangzhou: Zhejiang University, 2007.

[12] Rasiwasia N, Costa Pereira J, Coviello E, et al. A New Approach to Cross-modal Multimedia Retrieval [C]. In: Proceedings of the International Conference on Multimedia. ACM, 2010: 251-260.

[13] Costa Pereira J, Coviello E, Doyle G, et al. On the Role of Correlation and Abstraction in Cross-modal Multimedia Retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 521-535.

[14] Liu Yang, Zheng Fengbin, Jiang Baoqing, et al. Research of Cross-media Information Retrieval Model Based on Multimodal fusion and Temporal-spatial Context Semantic [J]. Journal of Computer Applications, 2009, 29(4): 1182-1187.

[15] Zhai X, Peng Y, Xiao J. Cross-modality Correlation Propagation for Cross-media Retrieval [C]. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan. IEEE, 2012: 2337-2340.

[16] Zhang Yu, Liu Yudong, Ji Zhao. Vector Similarity Measurement Method [J]. Acoustic Technology, 2009, 28(4): 532-536.

[17] (Reference for mathematical theories and derivations of partial least squares)

Author Contributions

Lu Wei: Designed the research framework, revised the final manuscript.
Ding Heng: Proposed the research proposition, designed the implementation plan, performed data analysis and processing, drafted the manuscript.

Received: 2015-07-06
Revised: 2015-09-16

Note: Figure translations are in progress. See original paper for figures.