Document Dependent Fusion in Multimodal Music Retrieval

Zhonghua Li, Bingjun Zhang, Ye Wang
School of Computing, National University of Singapore, Singapore
{lizhongh, bingjun, wangye}@comp.nus.edu.sg

ABSTRACT

In this paper, we propose a novel multimodal fusion framework, document dependent fusion (DDF), which derives the optimal combination strategy for each individual document in the fusion process. For each document, we derive a document weight vector by estimating the descriptive abilities of its different modalities. The document weight vector also enables our framework to be easily integrated with existing multimodal fusion schemes, and achieve a better combination strategy for each document given a query. Experiments are conducted on a 17174-song music database to compare the retrieval accuracy of traditional query independent fusion and query dependent fusion approaches, and that obtained after integrating DDF with them. Experimental results indicate that DDF can significantly improve the retrieval performance of current fusion approaches.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Query formulation; H.5.5 [**Sound and Music Computing**]: Systems

General Terms

Algorithms, Design, Experimentation

Keywords

Music, multimodal search, query dependent fusion, document dependent fusion, descriptive ability

1. INTRODUCTION

Since most documents (e.g., music, images, text documents) contain information or cues in different modalities, multimodal fusion, which aims to combine these modalities to better meet users' information needs, has been regarded as an effective approach to improve IR performance. Intensive research has been carried out aiming to estimate the optimal combination strategy of different modalities. Existing multimodal fusion approaches mainly include two categories: query independent fusion (QIF) and query dependent fusion (QDF). In the QIF scheme, the same fusion strategy is applied to combine different modalities for every query regardless of the query content and diversity. An example work of this category is CombSUM/CombMNZ method found in meta search [5]. However, this strategy is inadequate due to the fact that all queries are not equally created and they may have different information focuses. QDF emerges with the goal to improve the retrieval performance by adopting different fusion strategies for each query class (e.g., [1, 2, 7, 6]) or individual query (e.g., [9]). Noted that these approaches focus on deriving the optimal fusion strategy by only considering queries. Once the fusion strategy is determined (either in QIF or QDF), all the documents adopt the same strategy to combine its modalities.

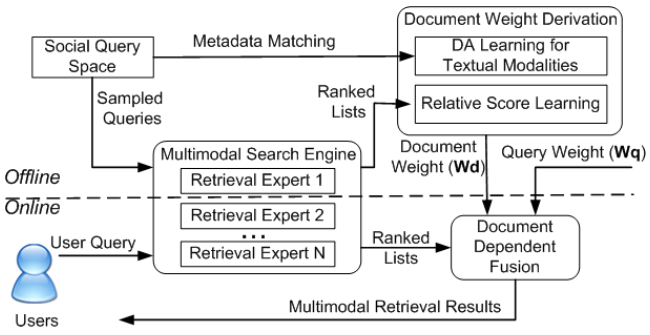

Figure 1: Document Dependent Fusion Framework for Multimodal Information Retrieval

In this paper, we propose a document dependent fusion (DDF) framework to derive the optimal fusion strategy for each document by considering influences of both queries and documents on the retrieval performance of IR applications. The structure of the framework is illustrated in Fig. 1. A document weight vector is derived for each document by learning the descriptive abilities of different modalities. This weight vector can be then integrated with QIF and QDF approaches. Confirmed by our experimental results, DDF significantly improves the performance of multimodal music retrieval.

The rest of the paper is organized as follows: Section 2 describes our approach from three main aspects: framework overview, document weight deviation, and document dependent fusion. Section 3 describes configurations of our experiments. Experimental results are presented and analyzed in Section 4, followed by the conclusion in Section 5.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

2. PROPOSED APPROACH

In this section, we first present an overview of the DDF framework based on the basic multimodal fusion structure. The fusion weight learning is then introduced in two phases: document weight deviation and document dependent fusion.

2.1 Document Dependent Fusion Overview

A multimodal retrieval system generally has N retrieval experts $\mathbf{RE} = [RE_1, ..., RE_N]$ to search relevant documents from different modalities. Given a query q , each retrieval expert RE_i returns a ranked list with the top n documents: $[d_{i1}, d_{i2}, ..., d_{in}]$. These ranked lists can be combined using a weight vector $\mathbf{W} = [W_1, ..., W_N]$, where $0 \leq W_i \leq 1$ and $\sum W_i = 1$. Each unique document d_k ($k \in [n, N * n]$) that appears in these ranked lists will be ranked by the final score

$$S_{d_k,q} = \sum_{i=1}^{N} s_{i_q,d_k} * W_i, \tag{1}$$

where s_{i_q,d_k} represents the score of document d_k in the ith ranked list given the query q, and it is generally normalized.

QIF approaches adopt the same fusion weight $\mathbf{W}$ for all the queries. The QDF scheme derives the optimal $\mathbf{W}$ for each query class or individual query. Both schemes provide a query-to-weight mapping in different resolutions. We format the fusion strategy in both cases as $\mathbf{W} = \mathbf{W_q}$. In this work, $\mathbf{W_q}$ and $\mathbf{W}$ are termed query weight and fusion weight.

We propose a document dependent fusion framework to derive the optimal fusion strategy for each document. A document weight vector ($\mathbf{W_d}$) is derived for each document according to the contributions of its different modalities. This weight vector can be integrated with the query weight to achieve a better fusion strategy (Eq. 2) for this document.

$$\mathbf{W} = f(\mathbf{W_q}, \mathbf{W_d}). \tag{2}$$

This fusion weight is used in the final score ranking in Eq. 1.

2.2 Document Weight Derivation

In this section, we derive a document weight vector which can reflect the importance of different modalities to a document. Since the formats of different modalities (e.g., textual, content) vary, we define a descriptive ability (DA) for each modality as a measure of its importance. Intuitively, a piece of information is considered useful only if it can meets users' information needs. The DA of a modality is defined as how well this modality meets users' information needs. We proceed to introduce the user information need space to lay the foundation of descriptive ability measures.

2.2.1 User Information Need Space

Since users' information needs are generally conveyed in queries, a large number of user queries need to be collected in order to construct a good user information need space. To save the effort of manually designing or collecting user queries, we construct a online folksonomy-based music social query space to simulate users' information needs in music retrieval.

The data information of this space is presented in Table 1 More detailed descriptions of this space can be found in [9].

2.2.2 Descriptive Ability of Textual Modalities by Metadata Matching

Textual metadata (e.g., title, descriptions) associated with music is an important modality of music documents. In multimodal music retrieval, there are generally several textual retrieval experts match music metadata with user queries on different music dimensions (e.g., genre, mood).

Table 1: Tag information in the music social query space. No. is the total number of tags in each music dimension

Dimensions	No.	Tag Clusters
Genre	244	Classical, Country, Electronic, HipHop, Jazz, Metal, Pop, Rock
Mood	286	Angry, Joy, Pleasure, Sad
Vocalness	13	Female, Male, Mixed, Nonvocal
Instrument	454	Brass, Percussion, Strings, Woodwinds

The metadata of each music document is compared with the music social query space. The binary occurrence count $n(d, w)$ of every word w from the music social query space (M) in document d is calculated. If word w appears in document d, $n(d, w)$ is set to 1, otherwise, we set it to 0. The descriptive ability of the kth textual modality of document d, can be derived as:

$$DA_T(d, k) = \frac{\sum_{w \in M_k} n(d, w)}{\sum_{w \in M} n(d, w)}, \tag{3}$$

where M_k is a subset of words in the music social query space M, and contains only words belong to the kth music dimension.

2.2.3 Descriptive Ability of Audio Modalities by Relative Score Learning

Audio modality refers to information derived from audio content, and is generally represented as feature vectors. Unlike metadata matching for textual modalities, these feature vectors are not directly comparable with the information needs represented using query keywords.

In this work, we utilize the relative score between audio and textual modalities to derive the descriptive ability of audio modalities. To estimate this relative score, we statistically analyze the retrieval results using a large number of user queries. For each query, N ranked lists are returned by all retrieval experts based on the similarity measure between the query and each modality of the documents. For each document d_k, we record the rank score s_{i_q,d_k} in the ith ranked list given a query q, and accumulate its total appearance over all the queries. The average score for the ith modality of this document can then be represented as:

$$S_{av}(i, d_k) = \sum_{q}^{N_Q} s_{i_q,d_k} / \sum_{q}^{N_Q} f_s(i_q, d_k), \tag{4}$$

where N_Q is the total number of queries; $f_s(i_q, d_k)$ is a selector function, which is set to 1(0) only when document d_k appears (does not appear) in the ith ranked list given a query q. We set $S_{av}(i, d_k)$ to 0 if document d_k never appears in the ith ranked list.

The average score provides an approximate importance of a modality over the tested queries. The relative score R_{j,d_k} between audio content and textual modalities on music dimension j ($j \in [1, N/2]$) is calculated as

$$R_{j,d_k} = \begin{cases} \frac{S_{av,C}(j,d_k)}{S_{av,T}(j,d_k)} & \text{if } S_{av,T}(j, d_k) \neq 0 \text{ and } S_{av,C}(j, d_k) \neq 0 \\ \kappa & \text{if } S_{av,T}(j, d_k) = 0 \text{ and } S_{av,C}(j, d_k) \neq 0 \\ 1/\kappa & \text{if } S_{av,T}(j, d_k) \neq 0 \text{ and } S_{av,C}(j, d_k) = 0 \\ 1 & \text{if } S_{av,T}(j, d_k) = 0 \text{ and } S_{av,C}(j, d_k) = 0, \end{cases} \tag{5}$$

where $S_{av,C}$ and $S_{av,T}$ are the average scores for content and textual modalities, respectively. Since the sampled queries

may not cover all the cases in the music social query space, we introduce κ to balance the cases when any of the average scores is zero. In our experiments, κ is set to 1. The descriptive ability for the jth audio modality of document d_k can be calculated as

$$DA_C(d_k, j) = R_{j,d_k} * DA_T(d_k, j). \quad (6)$$

2.2.4 Document Weight Learning

Given the descriptive abilities of a document, document weight vector is derived by assigning higher weights to modalities with larger descriptive abilities. Assume W_{d,T_j} (W_{d,C_j}) is the weight for the jth textual (audio content) modality of document d, the document weight vector is represented as

$$\mathbf{W_d} = [W_{d,T_1}, ..., W_{d,T_{N/2}}, W_{d,C_1}, ..., W_{d,C_{N/2}}],$$

where $W_{d,T_j} = \lambda DA_T(j)$ and $W_{d,C_j} = \lambda DA_C(j)$; λ is the normalization constant, which can be calculated by solving the following equation:

$$\sum_{j=1}^{N/2} (W_{d,T_j} + W_{d,C_j}) = 1. \quad (7)$$

2.3 Document Dependent Fusion

The final fusion weight $\mathbf{W}$ integrates both query dependency and document dependency. Given a query q, with the query weight vector $\mathbf{W_q} = [W_{q,1}, ...W_{q,N}]$ estimated using existing approaches, all the unique documents returned by a retrieval expert share the same query weight. Assume the document weight for document d is $\mathbf{W_d} = [W_{d,1}, ..., W_{d,N}]$. Its final fusion weight $\mathbf{W} = [W_1, ..., W_N]$ is computed as:

$$W_i = \frac{W_{d,i} * W_{q,i}}{\sum_{j=1}^{N} W_{d,j} * W_{q,j}}, \quad (8)$$

where $i \in [1, N]$. From Eq. 8 we note that previous fusion schemes, QIF and QDF, are special cases of DDF when the document weight vector is set to $\mathbf{1}$ for every document.

3. EXPERIMENTAL CONFIGURATION

3.1 Data Collection

Our database contains 17174 music tracks together with their metadata. The metadata includes title, descriptions, keywords, comments from YouTube and tags from Last.fm. For each audio track, the ground truth music styles in four music dimensions (genre, mood, vocalness, instrument) were annotated and cross checked by amateur musicians.

Totally 236973 social queries were generated from the music social query space following the guidelines in [9]. We randomly sampled 200K queries for relative score learning (Section 2.2.3) and query weight training, and the remaining ones served as testing data.

3.2 Retrieval Methods

We consider both textual and audio retrieval experts for each music dimension being studied. Each incoming query is parsed by comparing it with the music social query space. The discovered keywords are then fed to corresponding textual and audio retrieval experts.

For textual retrieval expert, we adopt the standard text retrieval approach: first tokenize the metadata and eliminate the stop words, followed by stemming using Porter's algorithm [3] and ranking using OKAPI BM-25 [4].

In audio retrieval processes, each audio track is represented as an audio signature, which is Fuzzy Music Semantic Vector (FMSV) [8] representing the probabilities that a music item belongs to different music styles. Given a query keyword, a query FMSV is generated by filling 1 to the matched music style and 0 to other music styles. Euclidean distance is calculated between the query FMSV and the ones of all the audio tracks. A final ranked list is constructed by ranking tracks with smaller Euclidean distances higher in the list.

The score s_{i,d_k} ($k \in [1, N_d]$) of document d_k in the ith ranked list is normalized as: $s_{i,d_k} = 1 - Rank_i(d_k)/n$, where N_d is the number of unique music documents retrieved by all retrieval experts, and n is the number of music documents in each ranked list. We set n to 100 in our experiment.

As multiple music dimensions are considered, we adopt fractional relevance score to address the partial match between a music document and a query. This relevance score is determined as the number of matched dimensions to the total number of dimensions required by the query.

3.3 Comparison Methodology

To evaluate the proposed framework, we compared the performance of existing fusion schemes (QIF and QDF) with DDF using average precision (AP) and mean average precision (MAP). The MAP differences among various approaches were further assessed using t-test.

In the QDF scheme, we implemented *Regression-based QDF (RQDF-ORPegasos)*, which predicts a fusion strategy for each query using a regression model based on oracle combination weights [9]. After integrating document weights with this method, we named it *DRQDF-ORPegasos*.

We also implemented two QIF approaches. The first approach applies an optimal weight combination to all the queries based on oracle combination weights [2]. The second approach treats all retrieval experts with equal combination weights. We named these two approaches as QIF and equally weighted QIF (*EQIF*), and after integrating with DDF they were termed *DQIF* and *DEQIF*, respectively.

For each query, grid search was applied to derive the oracle combination weight which produced the highest AP. The query and its oracle combination weight formed a training sample and were used in *RQDF-ORPegasos* and *QIF* approaches. We examined the performance when different sizes of training datasets were used. Given a query in the testing dataset, each fusion method was tested three times, and the average AP of these three tests and MAP were computed over all the testing queries.

4. RESULTS AND ANALYSIS

Fig. 2 presents the retrieval accuracy (MAP) improvement of DQDF approaches compared to QDF approaches. Fig. 3 depicts the MAP comparisons between DQIF and QIF approaches. Table 2 illustrates the detailed retrieval accuracy improvement in different query types and the t-test results. *RQDF-ORPegasos* adopted 5 training samples to compute the sub-gradient in each iteration of ORPegasos [9].

4.1 QDF vs DQDF

When comparing with *RQDF-ORPegasos* (Fig. 2, Table 2), our approach performs much better when the training data size is below 6.4K, and fails to compete with *RQDF-ORPegasos* when the training query set becomes larger. This might be attributed to the relative score learning between textual and audio modalities. Since the relative score is learned from a subset of queries randomly selected from the original query

Table 2: Retrieval accuracy (MAP) improvement in different query types when integrating document weights with previous approaches. G, M, V, I represent genre, mood, vocalness, and instrument four music dimensions, while A indicates all four dimensions. t-test was conducted for the overall improvement, and p values are attached in column A. "<" means $p < 0.05$, and ">" means $p > 0.05$. No.(*) represents the size of training dataset, $* = \times 10^3$. The content in each column means the improvement in percentage (%). "-" means the performance decreases after integrating with document dependency.

Method	No.(*)	A(p)	G	M	V	I	GM	GV	GI	MV	MI	VI	GMV	GMI	GVI	MVI	GMVI
DRQDF-ORPegasos	0.1	3.1 (<)	-0.1	1.6	1.1	-0.3	6.9	6.5	0.8	3.6	2.1	6.5	4.0	1.8	4.0	4.0	2.3
	1.6	1.5 (<)	-0.1	1.6	0.7	-0.3	3.4	2.6	0.8	1.0	2.2	2.7	0.3	2.0	2.3	1.7	0.8
	25.6	-0.8 (<)	0	1.5	0.7	-0.3	1.9	1.0	-0.1	0.2	0.6	1.4	-2.9	-1	0.3	-0.6	-2.3
	200	-0.6 (<)	0	1.5	0.7	-0.3	2.3	1.6	0.1	0.1	0.6	1.5	-2.5	-1	0.7	-0.8	-2.2
DQIF	0.1	0.5 (>)	-0.1	1.6	0.8	-0.3	11.8	13.3	0.3	2.6	0.3	1.1	1.4	-0.9	-0.7	0.3	-1.9
	1.6	0.3 (>)	-0.1	1.6	0.8	-0.3	11.5	13.6	0.1	3.2	0.4	1.5	1.8	-1.1	-1.1	0.2	-2.7
	25.6	0.2 (>)	-0.1	1.6	0.8	-0.3	11.5	13.5	0.1	3.1	0.3	1.4	1.7	-1.1	-1.2	0.2	-2.8
	200	0.2 (>)	-0.1	1.6	0.8	-0.3	11.4	13.5	0	3.1	0.3	1.4	1.6	-1.2	-1.2	0.2	-2.8
DEQIF	N.A.	10.1 (<)	0.3	1.4	2.2	-0.2	9.7	13.7	12.9	5.6	15.4	19.3	6.2	17.0	17.3	10.9	5.6

dataset, it only provides an approximation of the relative importance between different modalities. The number of queries used in this process and the distributions among different music dimensions may affect relative scores in representing the real descriptive abilities of different modalities.

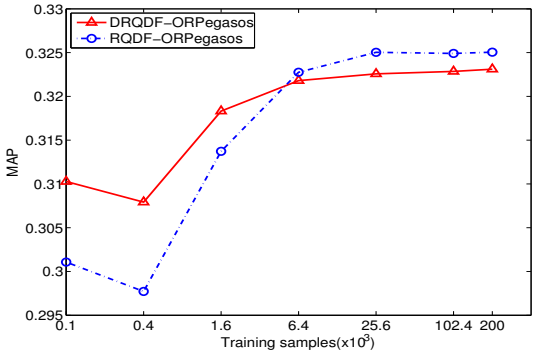

Figure 2: Retrieval accuracy (MAP) improvement when integrating DDF with RQDF-ORPegasos.

4.2 QIF vs DQIF

When comparing with QIF scheme, our approach outperforms QIF with an average improvement of 0.3% in MAP (Fig. 3). For *EQIF*, our approach increases the MAP by 10.1% (from 0.268 to 0.295). It indicates that the proposed framework unleashes the power of each document modality in multimodal fusion frameworks.

The MAP improvement in Table 2 illustrates the superior performance of our approach in estimating the optimal fusion weights in many query types. For instance, nearly all the 2-dimensional and 3-dimensional queries, and two single-dimensional (mood and vocalness) queries are improved using DDF approaches. The performance of query by genre and instrument is decreased with an almost constant ratio. This is mainly due to the insufficient number of queries in a particular query type while learning relative scores.

5. CONCLUSIONS

We have introduced a novel multimodal fusion framework, document dependent fusion (DDF), to derive the optimal fusion strategy for each individual document. We derive a document weight vector for each document, and integrate it into the fusion process. Confirmed by the experimental results, our proposed framework has significantly improved current multimodal fusion approaches. Moreover, in the process of DA derivation, relative scores of different modalities are learned from the retrieval results of a large number of

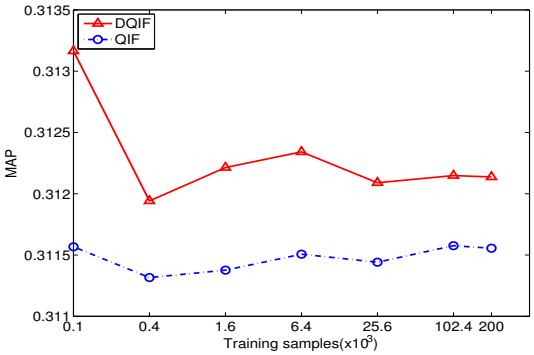

Figure 3: Retrieval accuracy (MAP) improvement when integrating DDF with QIF.

user queries. By combining with the DA of textual modalities, this method can derive DAs for modalities with various formats. Therefore, it can also be extended to other media documents in different multimodal retrieval applications, such as meta search, and video/image retrieval.

6. ACKNOWLEDGMENTS

The work was supported by Singaporean MOE grant R-252-000-381-112.

7. REFERENCES

[1] T. S. Chua, S. Y. Neo, K. Y. Li, G. Wang, R. Shi, M. Zhao, H. Xu, Q. Tian, S. Gao, and T. L. Nwe. Trecvid 2004 search and feature extraction task by nus pris. In *NIST TRECVID Workshop*, 2004.

[2] L. S. Kennedy, A. P. Natsev, and S. F. Chang. Automatic discovery of query-class-dependent models for multimodal search. In *Proc. of ACM Multimedia*, 2005.

[3] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.

[4] S. E. Rrobertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *TREC*, 1995.

[5] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *TREC-2*, 1994.

[6] L. Xie, A. Natsev, and J. Tesic. Dynamic multimodal fusion in video search. In *Proc. of IEEE ICME*, 2007.

[7] R. Yan and A. G. Hauptmann. Probabilistic latent query analysis for combining multiple retrieval sources. In *Proc. of ACM SIGIR*, 2006.

[8] B. Zhang, J. Shen, Q. Xiang, and Y. Wang. Compositemap: a novel framework for music similarity measure. In *Proc. of ACM SIGIR*, 2009.

[9] B. Zhang, Q. Xiang, H. Lu, J. Shen, and Y. Wang. Comprehensive query-dependent fusion using regression-on-folksonomies: a case study of multimodal music search. In *Proc. of ACM Multimedia*, 2009.