4-2011

Fusing heterogeneous modalities for video and image re-ranking

Hung-Khoon TAN

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Fusing Heterogeneous Modalities for Video and Image Re-ranking

Hung-Khoon Tan
Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
hktan@cs.cityu.edu.hk

Chong-Wah Ngo
Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
cwngo@cs.cityu.edu.hk

ABSTRACT

Multimedia documents in popular image and video sharing websites such as Flickr and Youtube are heterogeneous documents with diverse ways of representations and rich user-supplied information. In this paper, we investigate how the agreement among heterogeneous modalities can be exploited to guide data fusion. The problem of fusion is cast as the simultaneous mining of agreement from different modalities and adaptation of fusion weights to construct a fused graph from these modalities. An iterative framework based on agreement-fusion optimization is thus proposed. We plug in two well-known algorithms: random walk and semi-supervised learning to this framework to illustrate the idea of how agreement (conflict) is incorporated (compromised) in the case of uniform and adaptive fusion. Experimental results on web video and image re-ranking demonstrate that, by proper fusion strategy rather than simple linear fusion, performance improvement on search can generally be expected.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Retrieval models]

General Terms

Theory, Algorithms, Performance Experimentation

Keywords

Modality agreement, graph fusion, heterogeneous modality fusion, re-ranking

1. INTRODUCTION

Learning from multi-modality fusion has been an effective and yet a long standing issue for information search. Past research efforts, especially for text-based retrieval, generally indicate that retrieval performance by combining evidences from different modalities outperforms those employing only single modality [17]. Plausible explanation to the success story includes that different modalities are likely to agree on relevant documents, but there tends to be less agreement or overlap in terms of the non-relevant documents [15]. Casting this explanation to multimedia domain, nevertheless, does not always render a clear understanding of how different modalities should cooperate, and even overly simplifies the underlying difficulty of fusion. With experience from TRECVID search task [23] for example, fusing multiple search experts of different modalities does not always promise consistently desirable performance for different types of queries, especially in the case of combining multiple poor retrieval experts.

The issue of fusions could become even more challenging when revisiting the problem in the web domain for retrieving images or videos. With the convenient platform provided by social media websites, the explosive growth of data includes not only content itself, but also the contextual information, ranging from user tags, descriptions and peer comments. The difficulties of fusing the observations from various sources in the web domain are depicted in Figure 1.

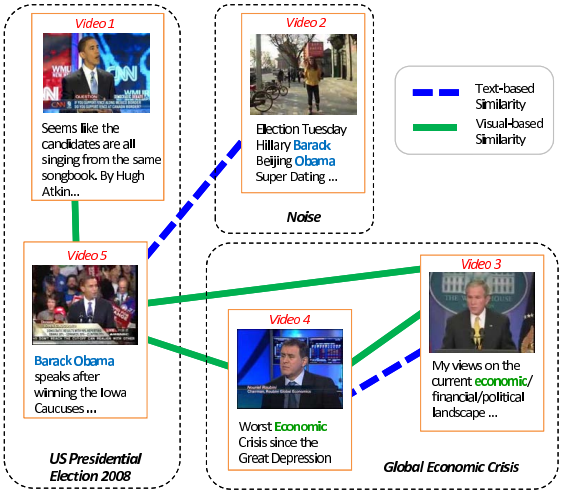
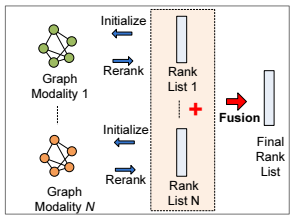
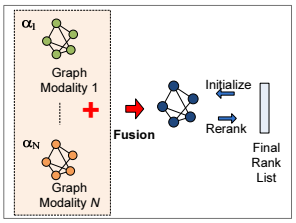


Figure 1: Five videos from YouTube for the query 'US Presidential Debate 2008'. The textual and visual similarity generate two sets of very different hyperlinks. The challenge is how to fuse these inconsistent heterogeneous information for ranking videos according to their relevancy.

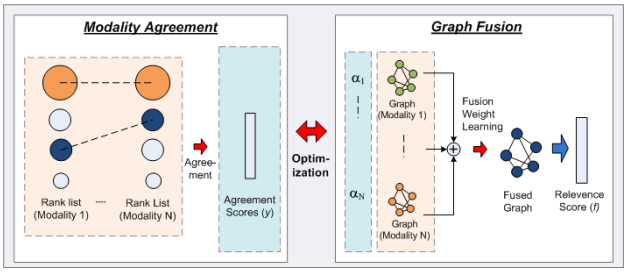
ICMR'11 April 17-20, Trento, Italy



(a) Search Expert Fusion



(b) Linear Graph Fusion



(c) Simultaneous optimization of Modality Agreement and Graph Fusion

Figure 2: Different multiple modality fusion techniques

First, user-supplied descriptions including tags tend to be subjective (video 1), incomplete (video 1) and noisy (video 2). A recent survey in [5] reported that, on the manual labeling of 81 semantic concepts for about 0.26 million Flickr images, about half of the user tags are noisy and half of the true labels are missing from user tags. Similar or even worse statistics is expected for web videos. Second, multimedia documents can be represented in diverse ways where different features can be extracted to represent their contents. The diversity is even more severe for video document whose content changes across time and includes audio content. Third, some modalities are conflicting with each other. Two images may be visually similar for having similar backdrop but contain different labels for foreground objects, or vice versa (video 3 and 5). Two videos of the same topic may be connected by overlapping near-duplicate clips but narrate different peripheral events. Though the videos are related topically and together they describe the evolution of a topic, the tags and descriptions supplied by web users could be very different (video 1 and 5). In short, none of these modalities are sufficiently competent when considered separately since each can only present a partial and unique view of an image or video document.

One possible solution, in the context of image or video re-ranking, is to emulate the success of PageRank [19] to fuse multiple information sources [13]. To extend PageRank to multi-modality fusion, multiple link graphs (one for each modality) can be constructed as in Figure 2(a). In search expert fusion (SEF), PageRank is conducted in parallel on each graph to produce multiple reranked lists, representing different views of search results from various modalities. The problem then becomes the late fusion of multiple search lists using different normalization techniques [9, 20] and combination operators [10]. More advanced techniques based on linear graph fusion (LGF) linearly combine the graphs of multiple modalities, as in Figure 2(b), where the fusion weights to generate a fused graph, if training samples are available [28]. Then, a single run of Random Walk [13] or semi-supervised learning [28] algorithm are conducted on the fused graph to produce the final rank list.

This paper addresses the fusion of heterogeneous modalities from the perspective of optimizing modality agreement while automatically predicting modality significance, two issues that are not adequately addressed by SEF and LGF. While the significance of modalities can be utilized to build a realistic fused graph for LGF, such knowledge is practically not available in general cases. Moreover, the importance of a modality could vary depending on the query topic as well as the search task at hand. Finding appropriate fusion weights for multiple modality graphs is thus not trivial, and is expected to impact the search results if the weights are not properly set. In addition, considering the conflicting nature of some modalities, not all search experts in SEF may agree with each other on the fused result, and it is difficult to resolve these conflicts with late fusion. SEF clearly lacks the interaction at the model level among the modalities, while in LGF, there is no mechanism to tell the degree of how different search experts (dis)agree on the final rank list.

We revisit both paradigms and cast the problem of fusion as a simultaneous mining of agreement from different modalities, while adapting fusion weights to generate a fused graph of multiple modalities to propagate the agreement. Figure 2(c) gives an overview of the proposed framework. Each search expert, corresponding to a certain modality, presents its own view of search result with a rank list. The initial agreement sought from these rank lists is utilized to initialize and construct a fused graph[1]. A new rank list is then produced by information propagation in the fused graph. The new list is in turn assessed by different modalities and new agreement is sought again to re-initialize the fused graph with new fusion weights. The agreement-fusion optimization iterates until convergence. The significance of our work lies in the capability to consider partial views from different modalities and predict modality importance by incorporating their agreement while compromising conflicts. Based on Figure 2(c), we adopt two well-known algorithms: random walk [13] and semi-supervised learning [28] for multi-modality fusion.

The remaining of the paper is organized as follows. Section 2 discusses related works on multi-modality fusion in web domain. Section 3 proposes a general framework for simultaneous agreement mining and graph fusion. We demonstrate the efficacy of the proposed algorithm, for video and image re-ranking in sections 4 and 5, respectively. Finally, section 6 summarizes the major findings and concludes this paper.

2. RELATED WORKS

Over the years, the growing number of heterogeneous data in the web has led to growing research interest into how to fuse the diverse information from different modalities. Fusion can be performed either at the periphery level [3, 11, 16, 30] without performing any drastic modification to the single modality framework, or at the model level [13, 18, 31, 8] where fusion modifies the model structures or formulation.

[1]Note that a set of modalities different from the search experts can be employed to build the fused graph. For example, search experts can be derived from the ranked lists of different search engines, and the fused graph from different content features or metadata.

The peripheral approaches consider data fusion as supplementary to existing models where it is performed either as early or late fusion. Early fusion [24] concatenates the features from different modalities to form a single feature vector. To reduce the 'curse of high dimensionality', canonical correlation analysis [4] has also been proposed to project the global feature into a lower dimensional subspace by correlating the data distribution in multiple views. Late fusion, on the other hand, comes into picture only after individual search experts produce their results. The scores can be combined in a linear [30, 16, 3] or non-linear [29] fashion. A previous work [25] has performed a comparative study between linear early and late fusion and concluded that late fusion consistently outperforms early fusion systems. The main problem with the peripheral approach is that there is limited interaction among modalities since these information are not employed to guide the actual search process.

The second category of approaches actively engages different modalities to modify the underlying model structures. In [13], a context graph similar to Figure 2(b) is created where the nodes represent image or video documents in the search set and the edges are based on the linear weighting between story-level text and visual duplicate similarities. However, the weights are pre-assumed and it is unknown how to properly determine the relative importance of the modalities. In [18], the PageRank model is extended to heterogeneous data where each link contains multiple parameters known as 'propagation factor' to model different types of relationship. The weights of each factor can be automatically learnt but it is based on the partial ranking of the objects given by domain experts. There has also been a lot of studies being done on multi-view clustering [22, 31] which performs simultaneous clustering of multiple graphs consisting of a common set of nodes. One natural approach is to convexly combine the kernels or affinity matrices for the graphs [27, 22]. Again, these approaches do not handle weight assignment for each kernel. To determine the optimal weights when combining graphs, [28] adopt semi-supervised learning using a regularization framework. Our proposed work, as shown in Figure 2(c), improves upon these approaches where the optimal weights are derived based on agreement. In addition, training samples are not required for initialization since by agreement, the set of pseudo positive documents agreed by most modalities can be automatically acquired.

3. AGREEMENT-BASED GRAPH FUSION

Given a set of P web images or videos, hereafter referred simply as documents, with M different modalities, the problem is to rank the relevance of these documents according to a given query topic. Assuming that each modality has an initial rank list of the documents, the problem becomes how to compromise the M initial rank lists and produce a final list that is satisfiable by all or most modalities. To model this problem, we depict the relationship among documents by M graphs $G_i = (V, E_i)$, where $i = 1, 2, \ldots, M$. The vertex set V common across all graphs represents the full set of P documents while the edge set E_i contains links between documents established based upon the distance function of ith modality. Each graph G_i is characterized by an affinity matrix $W_i \in \mathbb{R}^{P \times P}$, where each entry in W_i signifies the pairwise document affinity determined by E_i. To fuse M graphs from different modalities, we perform

$$\mathbf{W} = \sum_{i}^{M} \alpha_i W_i \tag{1}$$

where α_i is a fusion weight signifying the relevance of ith modality to a given query topic. Having the fused graph, the M initial rank lists can be projected to $\mathbf{W}$ with some combination operators. Algorithms such as semi-supervised learning and random walk could then be employed to re-rank the P documents.

There are two issues, nevertheless, regarding the aforementioned solution. First, the fusion weight α_i is query dependent and unknown in practice. Second, the M rank lists of documents may not agree with each other on the ranking scores f. Intuitively, fusion weights affect how the initial scores of documents are spread across the fused graph, while the agreement score represents the modality consensus on the relevance of documents. To deal with these two unknown variables, we seek for a new solution which simultaneously estimates the scores of documents based on the agreement among modalities and regulates the scores by graph fusion to produce a new rank list of documents.

3.1 Agreement Seeking

Given a query topic, each initial rank list of a modality basically presents a unique perspective of the document relevance to the topic and no single modality can adequately characterize the document-to-topic relevancy. Mining agreement from different partial views of modalities provides a more complete picture about the relevancy of documents. Denote the initial lists from M modalities as $S^{(0)} = \{s_1^{(0)}, ..., s_M^{(0)}\}$, where s_m is a P-dimensional vector containing the scores of P documents provided by the mth modality. The agreement score for a document i, denoted as $y(i)$, can be computed using different strategies depending on the types of provided rank lists as follows

Order-based Aggregation. Given a set of ordered lists (e.g., search list from different search engines), y can be derived from rank aggregation techniques such as Borda Count [1] and Markov Chain [9] aggregation. Similarly, we aggregate the inverse exponential function imposed on the ranking of each individual list as follows

$$y^{(0)}(i) = \sum_{m=1}^{M} e^{-\frac{R(s_m(i))^2}{0.02P}} \tag{2}$$

where $R(s_m(i))$ is the ranking index of document i in s_m. The constant 0.02 in the formulation is fixed empirically to ensure that the score drops rapidly with decreasing ranks in each list and thus aggregation assigns higher weight only to items highly ranked across multiple modalities.

Majority Voting. Some rankers additionally provide detection result (e.g. classification results from SVM models or detection results from pattern mining). In this scheme, only positive documents are allowed to place a vote and documents that accumulate at least ξ votes will be considered.

$$y^{(0)}(i) = \begin{cases} \sum_m^M s_m(i) & \text{if } |s_m(i)^+| \geq \xi \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $|\cdot|$ is the cardinality function and $|s_m(i)^+|$ is the number of modalities which labels i-th as positive.

Other possible alternative includes performing Mutual Information (MI) co-clustering [7, 12] which maximizes the

MI agreement when separating relevant from irrelevant set. Basically the first round of score assignment is based on the speculation that different modalities are more likely to agree on positive documents than negative documents [15]. Then, $y^{(t)}(i)$ will be updated over iterations and the negotiation power of M modalities on their contribution to score $y^{(t)}(i)$ may change, especially when the significances of modalities are revealed during graph fusion.

3.2 Graph Fusion

To fuse the graph, we can either plug in (a) semi-supervised learning or (b) random walk into the iterative framework to fuse multiple link graphs and propagate the agreement scores. Between the two, semi-supervised learning is superior in the sense that it can adaptively predict fusion weights as new agreement arrives at every iteration.

3.2.1 Semi Supervised Learning

Semi-supervised graph fusion is formulated as a regularization problem similar to [28], where the fusion weights in Equation 1 and the agreed document scores given by either Equation 2 or 3 are simultaneously optimized through EM algorithm. The graph-based learning enforces score smoothness in a way that similar documents should share similar scores. The score consistency is formulated using regularization as follows

$$f^* = \arg\min_{\mathbf{f}} \sum_{i=1}^{M} \alpha_i^r \Big(\sum_{j,k} W_i(j,k) \Big| \frac{f(j)}{\sqrt{D_i(j,j)}} - \frac{f(k)}{\sqrt{D_i(k,k)}} \Big|^2 + \mu \sum_{k=1}^{K} |f(p_k) - y(p_k)|^2\Big) \quad (4)$$

where $f(j)$ is the regularized score of the j-th document which we intend to find, $y(p_k)$ is its agreement score, $\{p_1, \ldots, p_K\}$ are the indices of the top-K documents in $f(j)$, D_i is the diagonal matrix for W_i and $\sum \alpha_i = 1$. The formula effectively propagates the scores from non zero-score documents to zero-score documents. The first term enforces label consistency where given two documents j and k, their scores $f(j)$ and $f(k)$ should be similar if the value in $W_i(j,k)$ is high so as to minimize the objective function. The second term is a regularization term that ensures f^* does not deviate too much from the initial score y. The tradeoff between the two terms is controlled by the damping factor $\mu > 0$. The exponent value r is introduced to control the degree of uniformity of α_i where as $r \to 1$, the bulk of the weight will be assigned to the single best modality whereas as $r \to \infty$, the values of α_i would be close to uniform.

A local solution to the minimization problem can be found through a variant of the EM algorithm. During the M-step, the value of fusion weight α_i is updated as follows

$$\alpha_i = \frac{1}{N_\alpha} (f^T L_i f + \mu \sum_{k=1}^{K} |f(p_k) - y(p_k)|)^{\frac{1}{1-r}} \quad (5)$$

where $L_i = I - D_i^{-\frac{1}{2}} W_i D_i^{-\frac{1}{2}}$ is the normalized Laplacian of W_i and N_α is the normalization factor so that the summation of all weights is equivalent to 1 or $\sum \alpha_i = 1$.

During the E-step, the value for f is updated using α_i from E-step as follows

$$f = (\mathbf{I} + \frac{1}{\mu N_f} \sum_{M}^{m=1} \alpha_m^r \mathbf{L}_m)^{-1} \quad (6)$$

where N_f is a normalization factor to ensure that $\sum \alpha_m^r = 1$. The E-step and M-step are iterated until convergence. Note that different from conventional semi-supervised learning such as [28], no manual labeling of training examples is required to initialize label y. Instead, y is the list of document scores given by Equation 4 based on modality agreement.

3.2.2 Random Walk

An alternative approach to exchange document scores is by PageRank like algorithm such as random walk [13]. Random walk on a graph can be viewed as a score distribution process by propagating the scores of documents to neighboring regions through edges of different weights. Naturally, documents with more neighbors are likely to receive more scores. To handle multiple graphs, a single context graph is created by linearly combining the affinity matrices of the graphs. Since to the best of our knowledge, there is no solution yet for random walk to learn the fusion weights, a uniform weighting scheme is assumed here. Random walk is then carried out as in the single graph case where the document score f is iteratively updated until convergence as follows

$$f = (\sum_i \alpha_i W_i) f + \mu y \quad (7)$$

and μ is a weighting factor which regulates how far the transition process is allowed to deviate from the 'personalization score' in the second term. The personalization vector is used to specify the preference of certain documents based on certain assumptions. In [13], it is derived from the text modality while in PageRank [19], a uniform distribution is used based on a random surfer model. In our case, the personalized vector is provided by the agreement scores obtained in Equation 2 or 3.

3.3 Heterogeneous Modality Fusion

Agreement modality and graph fusion could be conflicting. On one hand, the regulated document score f from graph fusion reflects the mutual relationship among documents on the fused graph. On the other hand, the agreement score y sought from multiple modalities only considers the relevance of each document individually. In face of two conflicting objectives, the agreement among modalities and the best possible fused graph might compete for dominance. To reconcile the difference, we relax the agreement factor through a damping term to achieve a balance between them through an iterative process. The agreement score $y^{(t)}(i)$ at iteration t is updated as follows

$$y^{(t)}(i) = \frac{y^{(t-1)}(i) + f^{(t-1)}(i) \times e^{\Delta t}}{N_y^{(t)}(i)} \quad (8)$$

where $N_y^{(t)}$ is the normalization factor so that $\sum_i y^{(t)}(i) = 1$ and Δ is the damping factor that controls the rate of convergence. As the iteration progresses, the document scores y will slowly converge to the second term. In the formulation, modality agreement will influence graph fusion through y whereas graph fusion in turn would influence modality agreement through the regularized scores f.

To avoid oscillation when optimizing both objectives, the damping factor Δ forces the two alternating factors to a compromise. This is synonymous to the annealing process where the temperature is gradually reduced as the simulation proceeds to allow the system settle harmoniously to a good low-energy region. In our case, the process effectively finds a local equilibrium between modality agreement and graph fusion. While the equilibrium state is not the optimal solution when viewed separately from either viewpoint, it represents a compromise between agreement and fusion. An optimal solution is reached when both factors are consistent. The algorithm is guaranteed to converge and the proof is a straightforward one since as$\Delta \rightarrow \infty$, the weights for the second term gains more prominence compared to the first term, i.e., $y^{(t+1)} \rightarrow f^{(t)}$. The algorithm is summarized in Algorithm 1.

Algorithm 1 Heterogeneous Modality Fusion

Initialization

1. Compute the initial agreement score $y^{(0)}$ (Eq. 2 or 3)
2. Set $t = 1$

Graph Fusion Step

(Semi-Supervised Learning)

1. Initialize $f^{(t)} = y^{(t-1)}$
2. E-Step: Update $\alpha^{(t)}$ (Eq. 5)
3. M-Step: Update $f^{(t)}$ (Eq. 6)
4. Repeat the E-step and M-step until convergence.

(Random Walk)

1. Perform a random walk process (Eq. 7)

Modality Agreement Step

1. Update the agreement score $y^{(t)}$ (Eq. 8)

Set $t = t + 1$ and repeat the graph fusion and modality agreement steps until convergence.

4. RE-RANKING YOUTUBE VIDEOS

In this section, we evaluate the proposed framework for re-ranking web videos. Given an initial list of videos, the task is to improvise the rank list by looking for consistencies among the videos.

4.1 Experiments and Evaluation

4.1.1 Experiment Setup

For evaluation, we use ten topics listed in Table 1. For each topic, we downloaded the set of videos from YouTube on May 2009 using multiple queries. The groundtruth are acquired through manual annotation where only typical videos directly related to the topic events are annotated as positive video. Some videos are somewhat related, and the guideline is to annotate videos that most users would like to see at the top of the search list. For example, for topic 2, the list contains a lot of user-generated videos in the form of dedication mtv for the shooting victims or 'video blogs' where users comment on the incident. We consider such videos as secondary and label them as negative. The re-ranking performance is measured with mean average precision (MAP) where the average precision (AP) for a topic is formulated as $AP = \frac{1}{N}\sum_{i=1}^{k} r_i/i$, N is the number of items we consider in the rank list, k is the number of positive items within the N items and r_i is the ranking of the i-th positive item. In our experiments, we set $N = 100$ for evaluation.

Table 1: Topic Dataset.

	Topics	#Vid	#Pos
1	US Presidential Election in 2008	737	145
2	The shooting event in Virginia Polytechnic Institute on 16/Apr/07	683	196
3	The Sichuan Earthquake on 12/May/08	1055	744
4	The forest fire disaster in California	426	92
5	Terrorist attack on London transportation network on 7th July 2005	784	150
6	The record high oil price crisis in 2008	759	318
7	Kosovo independence declaration in February 2008	524	417
8	Rusia presidential election in 2008	1335	116
9	Iran's nuclear enrichment program	1056	442
10	North Korea and weapon of mass destruction	1060	540

For our approach, we perform graph fusion through random walk using the agreement score and report the performance with and without iterations (AGRW and AGRW-IT, respectively). AGRW evaluates the performance of our algorithm gained purely from modality agreement, while AGRW-IT further evaluates the role of interaction between graph fusion and modality agreement towards the performance. The same evaluation is applied to the semi-supervised learning, denoted as AGSSL and AGSSL-IT. For the settings, the damping rate which controls the convergence rate is set to $\Delta = 0.3$. The fusion parameter which controls the degree of freedom for the modality weights is set to $r = 6$, while the minimum cross-modality overlap is set to $\xi = 3$. The top $K = 50$ pseudo labels from agreement are used to initialize graph fusion. Both the damping factor for random walk and regularization term in semi-supervised learning is set to $\mu = 0.75$.

Modalities. For the text modalities, we use the tag and descriptions (denoted as *tag* and *dsc*, respectively) associated with each video. After performing stemming and stop word removal, Apriori mining is applied to extract frequent tag and description keywordsets (denoted as $tagfks$ and $dscfks$, respectively). The minimum support count is set to 1% of the video count for the topic and only the 3- and 4-keywordsets are taken into account. The summation of support count of frequent keyword set is used as video scores. For the visual modality, near-duplicate segments are extracted using [26] to form near-duplicate threads, denoted as *ndt*. A thread is a group of near-dupicate video segments sharing similar scenes edited with extra captions or editing effects. In web videos, these segments are always used to connect videos of similar topics and can be used to reveal the evolution of a topic. The score of a video is based on the number of threads it is connected to. We use all five modalities (*tag*, *dsc*, $tagfks$, $dscfks$ and *ndt*) to construct five affinity graphs (W_i in Equation 1). We use only the scores from pattern mining ($tagfks$, $dscfks$ and *ndt*) for rank list in Equation 3, and thus the majority voting scheme is employed for agreement seeking.

Comparison. For comparison, we evaluate our algorithms against (a) Page Rank on individual modality (PR),

Table 2: MAP for the single modality versus heterogeneous modality fusion algorithms. The best result for each category is highlighted in bold. The overall best result for each topic is additionally marked with *.

Topic	1	2	3	4	5	6	7	8	9	10	MAP
PR (tag)	**0.246***	0.069	0.730	**0.086**	0.017	0.200	**0.616**	0.004	0.034	0.347	0.235
PR (dsc)	0.109	0.015	0.847	0.075	0.147	0.269	0.481	0.000	0.123	0.289	0.235
PR ($tagfks$)	0.057	**0.089**	0.799	0.076	0.056	**0.365***	0.488	0.013	0.134	0.453	0.253
PR ($dscfks$)	0.058	0.058	0.665	0.080	0.062	0.284	0.479	0.022	0.206	0.448	0.236
PR (ndt)	0.223	0.039	**0.860**	0.040	**0.170**	0.130	0.545	**0.088***	**0.295**	**0.586**	**0.298**
SEF	**0.097**	**0.111**	0.770	0.083	0.041	0.297	**0.492**	0.000	**0.305**	0.649	0.284
LGF	0.094	0.098	**0.784**	0.094	**0.084**	**0.371**	0.487	**0.030**	0.258	**0.709**	**0.301**
AGRW	0.148	0.130	0.812	0.086	0.126	**0.306**	0.520	0.040	0.257	0.695	0.312
AGSSL	0.119	0.118	0.817	0.078	0.129	0.297	0.554	0.021	0.273	0.724	0.313
AGRW-IT	0.158	0.130	0.804	0.083	0.120	0.306	0.526	0.048	0.262	0.696	0.313
AGSSL-IT	**0.177**	**0.150***	**0.877***	**0.108***	**0.281***	0.264	**0.668***	**0.084**	**0.404***	**0.790***	**0.380***

(b) search expert fusion (SEF) and (c) linear graph fusion (LGF) methods. For SEF, separate PageRanks are conducted on each modality graph. The set of rank lists are then normalized using Broda transformation which performs rank-based normalization as follows: $score(video) = N - rank(video)$ where N is the number of videos in the list. The normalized lists are then fused using CombSum [10] where the rank lists are linearly summed. For LGF, we perform linear graph fusion where a uniform weight is assumed and a single PageRank is run on the resultant graph.

4.1.2 Results and Analysis

Table 2 shows the MAP performance of the evaluated algorithms for the ten topics. In average, fusing multiple modalities does not always guarantee performance improvement over single modality. Compared to the best performing modality ndt, SEF causes a slight drop in MAP from 0.298 to 0.284 while LGF only manages to maintain roughly a similar performance at 0.301. This shows that mixing modalities of varying effectiveness risks diluting the performance of the more competent modalities. The performances of fusion are limited since they neglect the relative importance of the modalities. The best modality for PR in single modalities fluctuates wildly from topic to topic, as highlighted in bold in Table 2 although generally, the modalities generated by pattern mining ($tagfks$, $dscfks$ and ndt) are more reliable compared to the others (dsc and tag). This unpredictability of individual modality performance creates great uncertainty when fusing graphs in a uniform manner. In addition, the graph-level fusion is observed to be more robust compared to search expert fusion. This is because SEF only post-processes the rank list from each modality and does not involve any real interaction among the modalities, as opposed to LGF which uses a fused structure to carry out re-ranking.

With agreement, the MAP performance of graph fusion improves from 0.301 (LGF) to 0.312 (4%) for AGRW where AGRW either outperforms (topics 1, 2, 3, 5, 7 and 8) or maintain a similar performance to LGF. Similar to LGF, AGRW performs random walk on linearly fused graph, but additionally provides the agreement scores to strengthen the personalization distribution for PageRank. The improvement comes from the ability to perform an informed estimate of such preferences which makes the search become more focus and directed.

When the fusion weights are taken into account through semi-supervised learning, the MAP for AGSSL remains similar to AGRW (0.313). The improvement is limited because of non-ideal pseudo labels and this is observed indirectly by a more uniform weight assignment for all topics at the beginning of the iterations. The biggest improvement comes from iteratively optimizing graph fusion and modality agreement to achieve a balance between them. AGSSL-IT improves the MAP to 0.380, which is 21% and 26% better compared to AGSSL and LGF, respectively. On the other hand, no improvement is observed for AGRW-IT due to the lack of weight adaptation capability in the random walk framework. In fact, AGSSL-IT has the best MAP among all the compared algorithms on seven out of ten topics. Graph fusion and modality agreement compliments each other, where agreement is able to guide fusion in adjusting the modality weights, while fusion helps agreement to regularize their scores. However, the process may break down if fusion propagates more scores to videos which do not agree across modalities where a drop in performance is observed for topic 6 ($0.371 \rightarrow 0.306$). For majority of the topics (topics 4 to 10), the bulk of the final weights found by AGSSL-IT (ranges from 0.64 to 0.72) are assigned to the near-duplicate graph ndt. This is rather surprising since the result is not consistent with the best single modality (PR) in Table 2. The reason is because the framework assigns higher weight to the modalities, in our case ndt, which find better agreement with other modalities since optimization is guided by the scores acquired from agreement.

Discussion. From the experiment, we can make the following conclusions on the fusion of multiple heterogeneous modalities. First, the performance of uniform fusion schemes (SEF and LGF) is basically unstable because they tend to undermine the effectiveness of the more competent modalities. Re-ranking improves steadily as more information are shared among modalities from SEF (peripheral-level fusion) to LGF (model-level fusion) to AGRW (agreement) to AGSSL (agreement + weighting) to AGSSL-IT (agreement + weighting + iterations). Second, modality fusion is stable when the modalities are well-balanced, or in other words, no single modality overwhelmingly outperforms the rest. Consider topics 5 and 8 where ndt is the dominant modality, the MAP for SEF and LGF is significantly lower than single modality (PR). Indeed, it is critical to be able to apply the appropriate weighting in graph fusion as demonstrated by the superior performance of AGSSL-IT in the two topics.

Parameter Sensitivity. To evaluate the sensitivity of AGSSL-IT towards various parameters, we fix the damping factor Δ at various points ranging from 0.05 to 0.3 and then vary the value of the fusion parameter r from 1.1 to 20.

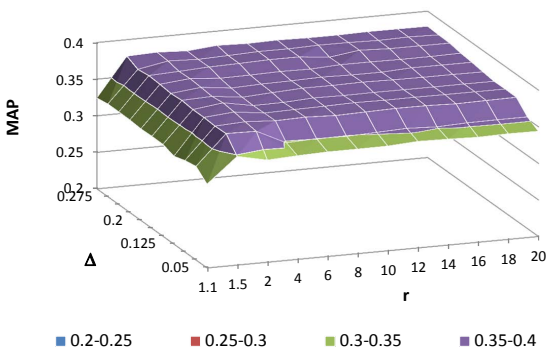


Figure 3: MAP result for different r and Δ.

A high value of r will enforce uniform weights for graph fusion, and vice versa. A low value of Δ indicates that the agreement scores are diffused slowly into graph fusion, and vice versa. Figure 3 shows the average MAP scores of all topics. In general, the algorithm is relatively stable where MAP stays within 0.37 to 0.38 for a wide range of settings. The best performance is achieved for $r = 6$ and $\Delta = 0.3$ and the stable operational range is from 2 to 20 for r and from 0.1 to 0.3 forΔ . MAP drops when $r < 2$ as the weight falls only on a single modality, and whenΔ < 0.1 where the algorithm is observed to converge to a less optimal local minima when the score diffusion is too slow.

5. RE-RANKING FLICKR IMAGES

In this section, we evaluate our approach for re-ranking web images on the recently released web image dataset, NUS-WIDE [6]. The corpus contains 269,648 images crawled from Flickr, containing a total 5,018 unique tags out of which 81 were manually annotated with clean ground-truths. The set is divided into two sets: 161,789 images for training, and 107,859 images for testing. For evaluation, the same settings as in section 4.1.1 are used for this experiment. We perform re-ranking on the top 1000 images retrieved by VIREO-WEB81 [32] which are SVM classifiers trained on the training set using Bag-of-Words (BoW) features with soft-weighting.

Modalities. For agreement, two rank lists are used. The first rank list is retrieved from VIREO-WEB81 (visual feature). The second list is the FCS rank list which is derived from textual features. In the FCS rank list, the scores for an image is given by summing the four highest Flickr Context Similarity (FCS) [14] scores between the query concept and the tags associated with the image. Since the score distributions of the FCS and VIREO-WEB81 rank lists are very different, modality agreement is performed by using Order-based Aggregation (Equation 2). For graph fusion, we use a different set of modalities for generating the affinity graphs W_i. We consider four modalities which are constructed based on the cosine distance on the (a) wavelet, (b) color moment, (c) BoW visual features for visual features and (d) the term frequency vectors from social tags associated with each image.

Results and Analysis. The experimental results are shown in Figure 4. The base MAP scores given by VIREO-WEB81 is 0.41 and FCS reranking is 0.51. By modality agreement (AG) between the two rank lists alone as elaborated in Section 3.1, the improvement is limited (MAP=0.53) since while modality agreement is efficacious for pinpointing relevant images, it is less certain when it comes to disagreeing ones (not necessarily negative image). When integrated with graph fusion (AGSSL), MAP jumps by 18% to 0.60 when video relationship are further taken into account through the fused graph. Modality agreement can only retrieve a certain portion of positive videos. As shown in Figure 5, graph fusion fills the gap by propagating the scores from agreement to the other positive images through the fused graph. In general, the results show that the modalities are quite balanced with a uniform weight scores with a slight preference to the tag link graph. This conforms with our baseline results which show that textual-based rank list FCS is the better than the visual-based rank list VIREO-WEB81. Iterating modality agreement and graph fusion (AGSSL-IT) improves the MAP only slightly to 0.62. One reason is because the initial solution found during the first iteration is already sufficiently optimal. Indeed, the fusion weight is observed to remain stable through the iterations for most concepts. This is different from re-ranking web videos in section 4 which becomes increasingly leaning towards the near-duplicate modality as the iteration progresses. More importantly, consistent improvement is observed for almost all concepts except for four concepts (nighttime, grass, town and window) with a minor decrease of around 5%. Upon investigation, the decrease in performance for the four concepts are mainly due to erroneous ground truth labels, mainly from false negatives.

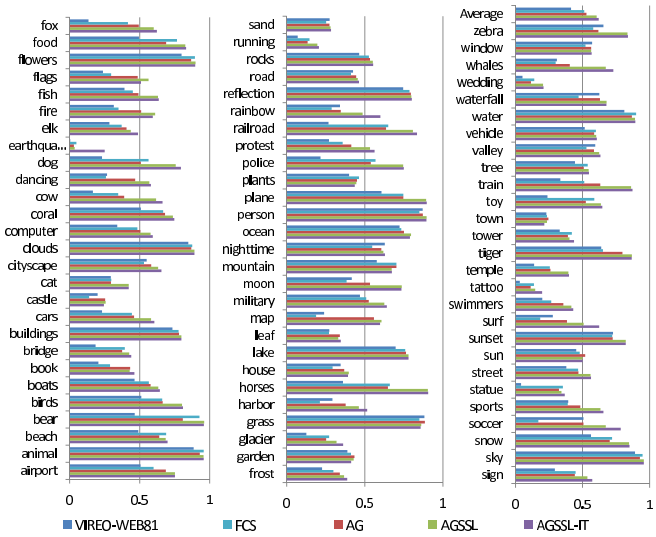


Figure 4: Re-ranking result for 81 concepts on NUSWide.

6. CONCLUSIONS

We have presented an agreement-fusion optimization model for fusing multiple heterogeneous data. The agreement between the scores from multiple modalities is explored to guide the fusion of multiple graphs in both linear and adaptive manners. The agreement is exploited in two ways, namely as the personalization distribution for random walk, or as pseudo training samples for semi-supervised learning to adapt the fusion weights of different modalities. To rec-

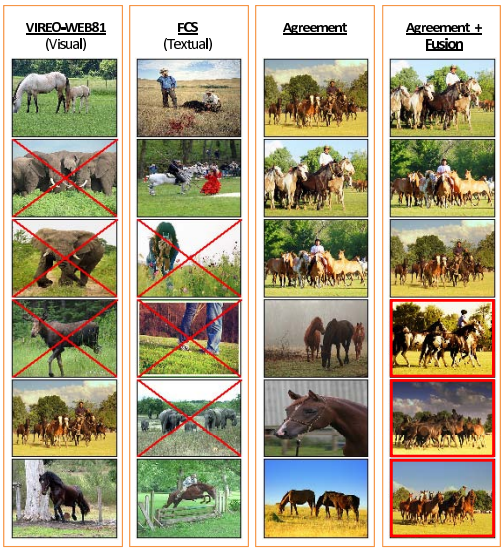


Figure 5: Top 6 re-ranking results for 'Horse'. Column 3 shows the result of seeking agreement on VIREO-WEB81 (column 1) and FCS (column 2) base rankers. Graph fusion (column 4) further pulls additional images (bottom 3 images) originally missed by agreement to the top of the list.

oncile the conflicting objectives between graph fusion and agreement, score exchange is conducted iteratively between the two steps to reach an equilibrium solution. We have also experimented the framework for reranking web images and videos. The experimental result reveals that different modalities perform wildly on different topics. Fusing multiple modalities generally lead to better performance than single modality. Nonetheless, without careful fusion strategy, linear fusion, such as SEF and LFG only shows marginal or worse improvement if compared to the single best modality (near-duplicate thread) in our experiment. By incorporating agreement from different modalities and further iterating agreement and fusion with dynamic weight adjustment leads to the best overall performance, which is achieved by the proposed AGSSL-IT. Our analysis indicates that the improvement comes from the adaptive weighting scheme based on the agreement of heterogeneous modalities. In the future, we would like to diversify our set of modalities for videos to include audio, face detector [21] and name-entity [2]. In addition, we would also like to extend the framework beyond reranking to a wider range of applications such video recommendation and summarization.

7. ACKNOWLEDGMENTS

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119508 and CityU 119610).

8. REFERENCES

[1] J. A. Aslam and M. Montague. Models for metasearch. *SIGIR*, 2001.

[2] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: A high-performance learning name-finder. *ANLP*, 1997.

[3] S.-F. Chang, J. He, Y.-G. Jiang, E. E. Khoury, C.-W. Ngo, A. Yanagawa, and E. Zavesky. Columbia university/vireo-cityu/irit trecvid 2008 high-level feature extraction and interactive video search. *TRECVID*, 2008.

[4] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. *ICML*, 2009.

[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. *CIVR*, 2009.

[6] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. *CIVR*, 2009.

[7] C. O. Conaire and N. O. Connor. Unsupervised feature selection for detection using mutual information thresholding. *WIAMIS*, 2008.

[8] H. Deng, M. R. Lyu, and I. King. A generalized co-hits algorithm and its application to bipartite graphs. *KDD*, 2009.

[9] C. DWork, R. Kumar, M. Naor, and D. Sivakumar. Rank agregation methods for the web. *WWW10*, 2001.

[10] E. A. Fox and J. A. Shaw. Combination of multiple searches. *Text REtrieval Conference*, 1994.

[11] J. Gao, W. Fan, Y. Sun, and J. Han. Heterogeneous source consensus learning via decision propagation and negotiation. *KDD*, 2009.

[12] G. Greco, A. Guzzo, and L. Pontieri. Co-clustering multiple heterogeneous domains: Linear combinations and agreements. *TKDE*, Nov 2009.

[13] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. *ACM Multimedia*, 2007.

[14] Y. G. Jiang, C. W. Ngo, and S. F. Chang. Semantic context transfer across heterogeneous sources for domain adaptive video search. *ACM Multimedia*, 2009.

[15] J. H. Lee. Analyses of multiple evidence combination. *SIGIR*, 1997.

[16] Y. Liang, B. Cao, J. Li, C. Zhu, Y. Zhang, C. Tan, G. Chen, C. Sun, J. Yuan, M. Xu, and B. Zhang. Thu-img at trecvid 2009. *TRECVID*, 2009.

[17] M. McCabe, A. Chowdhury, D. Grossmand, and O. Frieder. System fusion for improving performance in information retrieval systems. *ITCC*, 2001.

[18] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object level ranking: Bringing order to web objects. *WWW*, 2005.

[19] L. Page, S. Brin, L. Motwani, and T. Wingrad. The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*, 1998.

[20] M. E. Renda and U. Straccia. Web metasearch: Rank vs. score based rank aggregation methods. *SAC*, 2003.

[21] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *IJCV*, 56(3):151–177, 2002.

[22] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. *Text REtrieval Conference*, 1994.

[23] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. *MIR*, 2006.

[24] C. Snoek, M. Worring, J. Geusebroek, D. Koelma, and F. Seinstra. The mediamill trecvid 2004 semantic video search engine. *TRECVID*, 2004.

[25] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. *ACM Multimedia*, 2005.

[26] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua. Scalable detection of partial near-duplicate videos by visual temporal consistency. *ACM Multimedia*, 2009.

[27] H. Tong, J. R. He, M. J. Li, C. S. Zhand, and W. Y. Ma. Graph-based multi-modality learning. *ACM Multimedia*, 2005.

[28] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai. Optimizing multi-graph learning towards a unified video annotation scheme. *ACM Multimedia*, 2007.

[29] Y. Wu, E. Y. Chang, K. C. C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. *ACM Multimedia*, 2004.

[30] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. *SIGIR*, 2005.

[31] D. Zhou and C. J. Burges. Spectral clustering and transductive learning with multiple views. *ICML*, 2007.

[32] S. A. Zhu, G. Wang, C. W. Ngo, and Y. G. Jiang. On the sampling of web images for learning visual concept classifiers. *CIVR*, 2010. (http://vireo.cs.cityu.edu.hk/vireoweb81/).