Explicit and Implicit Concept-based Video Retrieval with Bipartite Graph Propagation Model

Lei Bao[1,2,3], Juan Cao[1], Yongdong Zhang[1], Jintao Li[1], Ming-yu Chen[3], Alexander G. Hauptmann[3]
[1]Laboratory for Advanced Computing Technology Research, ICT, CAS, Beijing 100190, China
[2]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[3]Graduate University of Chinese Academy of Sciences, Beijing 100049, China
{baolei, caojuan, zhyd, jtli}@ict.ac.cn, {alex, mychen}@cs.cmu.edu

ABSTRACT

The major scientific problem for content-based video retrieval is the semantic gap. Generally speaking, there are two appropriate ways to bridge the semantic gap: the first one is from human perspective (top-down) and the other one is from computer perspective (bottom-up). The top-down method defines a concept lexicon from human perspective, trains the detector for each concept based on supervised learning, and then indexes the corpus with concept detectors. Since each concept has an explicit semantic meaning, we name this kind concept as an explicit concept. The bottom-up approach directly discovers the underlying latent topics from video corpus by machine perspective using an unsupervised learning. The video corpus then is indexed by these latent topics. As opposite to explicit concepts, we name latent topics as implicit concepts. Given the explicit concept set is pre-defined and independent of the corpus, it is impossible to completely describe corpus and users' queries. On the other hand, the implicit concepts are dynamic and dependent on the corpus, which is able to fully describe corpus and users' queries. Therefore, combining explicit and implicit concepts could be a promising way to bridge the semantic gap effectively. In this paper, a **B**ipartite **G**raph **P**ropagation **M**odel (**BGPM**) is applied to automatically balance influences from explicit and implicit concepts. Concept nodes with strong connections to queries are reinforced no matter explicit or implicit. Demonstrated by the experiments on TREVID 2008 video dataset, BGPM successfully fuses explicit and implicit concepts to achieve a significant improvement on 48 search tasks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process.*

General Terms

Algorithms, Experimentation

Keywords

Semantic gap, explicit concept, implicit concept, bipartite graph

1. INTRODUCTION

Nowadays, with the explosive growth of video data, the demands for accessing to large video corpus are hardly served by present video retrieval applications. The major scientific problem for content-based video retrieval is the semantic gap between the visual low-level features and the high-level semantics [1].

The most popular approach to bridge the semantic gap is concept-based video retrieval [2]. It defines a concept lexicon from human perspective, trains a detector for each concept based on supervised learning, and then automatically indexes the video content with concept detectors. In the retrieval process, query is mapped to the concept set and then video clips which are the most relevant to concepts are returned to users. By using a simulation study, [3] shows that, "using no more than 5000 concepts will be sufficient for accurate retrieval to approach standard WWW search quality, despite a fairly low detection accuracy of 10% for any single concept and substantial combination errors". All of these indicate that concept-based video retrieval is a promising approach to bridge the semantic gap. However, due to the expensiveness of human labeling and the high computational cost for supervised learning, the available detectors are very limited right now. Furthermore, since the lexicon is pre-defined and independent of video corpus. Therefore, it is impossible to completely describe different kinds of corpus and users' queries with such a limited concept detector set, which limits the performance of retrieval.

Actually, the semantic lexicon is defined from human perspective. Therefore, the traditional concept-based retrieval could be considered as a top-down approach. As an opposition, we also could bridge the semantic gap by a bottom-up approach. Using unsupervised learning, this approach discoveries the underlying latent semantic structure (latent topics) of video corpus from machine perspective, and then indexes the video corpus by these latent topics. As the latent topics are dynamic and dependent on the corpus, they are able to fully describe corpus and users' queries. Many unsupervised learning methods are proposed for latent semantic structure discovery, such as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (PLSA), and Latent Dirichlet Allocation (LDA) [6]. Although, most of them are well applied to text retrieval, the related works in video domain are few. [7] and [8] successfully utilize LSI and LDA to find the semantics of visual features, thus help to improve the retrieval performance. These works indicate the potential of latent topics to bridge the semantic gap.

The essential difference between the top-down and bottom-up approaches is the way to define and discover semantic concepts. In the top-down approach, as each concept has an explicit semantic meaning from human perspective, we name this kind concept as an explicit concept. Meanwhile, in the bottom-up approach, as the latent topics are extracted from machine perspective, we name latent topics as implicit concepts.

According to above analysis on the top-down and bottom-up approaches, we could expect that combining explicit and implicit

Conference'10, Month 1–2, 2010, City, State, Country.
Copyright 2010 ACM 1-58113-000-0/00/0010…$10.00.

concepts is a promising way to bridge the semantic gap effectively. However, given that they are mutual complementary and their performance are various for different query tasks, a simple fusion is hard to believe to work [9]. In this paper, we apply the Bipartite Graph Propagation Model (BGPM) to automatically balance influences from explicit and implicit concepts. Firstly, the relationship between queries and the explicit and implicit concepts are modeled as a bipartite graph. Then concept nodes with strong connections to queries are reinforced by propagation. Therefore, the most important concepts will be highlighted no matter explicit or implicit. As demonstrated by the experiments on TREVID 2008 video dataset, BGPM successfully fuse explicit and implicit concepts to achieve a significant improvement on 48 search tasks.

The rest of this paper is organized as follows: Section 2 gives a comparison between the explicit and implicit concepts. Section 3 details the BGPM. Experimental results are provided in Section 4. Finally, the paper is concluded in Section 5.

2. EXPLICIT V.S. IMPLICIT

As we known, the explicit concepts detectors are predefined from human perspective and trained by supervised or semi-supervised learning, their performances could be evaluated by human labeling. Therefore, people could make efforts to improve the explicit concept detectors and finally improve the video retrieval. However, the implicit concepts are discovered by unsupervised learning from machine perspective. The evaluation for performance of the implicit concepts detectors is unavailable. As a result, we don't have a way to optimize implicit concept detectors to boost video retrieval result. This indicates if users' query could match a suitable concept from explicit concept set, the explicit concept-based retrieval will outperform the implicit concepts-based one. Unfortunately, the explicit concept set is predefined and independent of the video corpus. It is impossible to describe different kinds of corpus and queries. On the other hand, the implicit concept set is discovered from the underlying structure of the corpus. They are dynamic and dependent on the corpus, which is able to fully describe corpus and users' queries.

To further verify the above analysis, we designed a comparison experiment to observe the differences between the explicit and implicit concepts. For video set, we choose the TRECVID 2005 development set (TV05 in short) and TRECVID 2008 testing set (TV08 in short). As to the concept set, the explicit set is the CU-VIREO374 [10] which consists of 374 concepts predefined by LSCOM; the implicit set contains 80 concepts which are dynamically generated from TV08 by LDA [8]. The frequencies of the explicit and implicit concept sets on TV05 and TV08 are showed in Figure 1.The green and red lines separately denote the frequencies of CU-VIREO374 explicit concepts set in TV05 and TV08. The blue one denotes the frequency of the 80 implicit concepts in TV08. (1) As the green and red lines showed, TV05 covers well the whole explicit concept set, but TV08 only covers less half of it. Actually, most of videos in TV08 are documentary videos. However, most of videos in TV05 are news videos. It indicates that the predefined explicit concept set CU-VIREO374 fails to represent the different kinds of video corpus. It could work well in TV05 but not in TV08. (2) On the other hand, as showed by the blue line, the 80 implicit concept set gets a well cover in TV08. Most of them appear more than 100 times. It further verifies that the implicit concept set could completely describe the corpus, when the explicit concept set fails. At the time, resorting to implicit concept detectors could a wise choice.

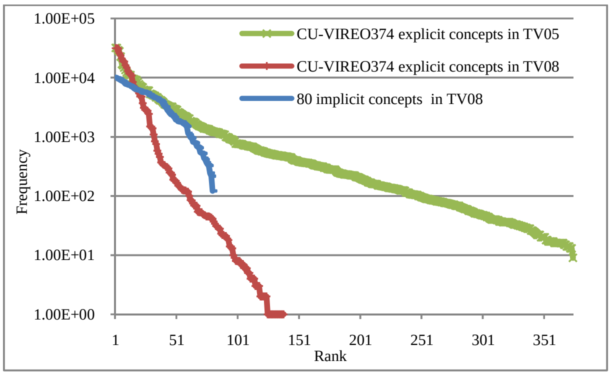

Figure 1. Frequencies of explicit and implicit concepts on different corpus (Y-axis is plotted in log scale).

All of these show that combining explicit and implicit concepts could be a promising approach to bridge the semantic gap effectively. However, for different kinds of video corpus and query task, the potential of explicit and implicit concepts are various. That inspires us to find a solution to automatically balance influences between explicit and implicit concepts.

3. BIPARTITE GRAPH PROPAGATION MODEL

In this paper, we implement Bipartite Graph Propagation Model (BGPM) [5] to automatically balance influences from the explicit and implicit concepts. Firstly, it is an intuitive idea to model the relationship between concepts and query-by-keywords or query-by-example as a bipartite graph, where one side is concept node set, the other side is query node set, and the relationship between queries and concepts are represented by the weighted edges between the two node sets. Secondly, according to the regularization of BGPM, concept nodes with stronger connections to query nodes will win after propagation stability. Since that, the most important concept no matter explicit or implicit will be automatically reinforced by BGPM. The predicted score of each concept node indicates its relevance to the queries. Finally, we use the predicted score as the weight for each concept, and the video clips are ranked according to the accumulated sum of weighted concept detector scores. In addition, besides fusion of the explicit and implicit concepts, another benefit brought from BGPM is that it unifies query-by-keywords and query-by-example, which makes it flexible in multi-modal query task.

3.1 Bipartite Graph Construction

The explicit and implicit concepts, query-by-keyword, query-by-example and their relationship are modeled in a bipartite graph $G=(Q,C,E,\mathbf{W})$, where Q and C are the concept node set and query node set, E is the edge set and $\mathbf{W}$ is the graph affinity matrix. In query node set $Q=\{q_1,q_2,\cdots,q_M\}$, q_1 denotes the query-by-keyword node and $\{q_2,\cdots,q_M\}$ represents the $M-1$ query examples. The explicit and implicit concepts are treated equally in the graph, where each node c_i in $C=\{c_1,c_2,\cdots,c_N\}$ indicates a concept and i is the index. $edge(i,j)\in E$ is assigned an affinity score w_{ij} to construct the $M\times N$ matrix $\mathbf{W}$ which reflects the relationship between query q_i and concept c_j . For w_{1j} , if the query-by-keyword contains a concept j , w_{1j} is set to

1, otherwise to 0. Obviously, for implicit concept nodes, it is 0 as well. For $w_{ij}(1 < i \le M)$, we use the concept detector score of concept c_j on query example q_i to represent the relationship.

3.2 Bipartite Graph Propagation

As shown in [5], the propagation process in the bipartite graph could be written as:

$$\begin{cases} \mathbf{f}_{t+1}^{c} = \alpha \widetilde{\mathbf{W}}^{T} \mathbf{f}_{t}^{q} + (1-\alpha)\mathbf{y}^{c} \\ \mathbf{f}_{t+1}^{q} = \alpha \widetilde{\mathbf{W}} \mathbf{f}_{t+1}^{c} + (1-\alpha)\mathbf{y}^{q} \end{cases},$$

where $\mathbf{f}_t^c = [f_1, f_2, \cdots, f_N]^T \in \mathbb{R}^N$ and $\mathbf{f}_t^q = [f_1, f_2, \cdots, f_M]^T \in \mathbb{R}^M$ are the updated score for concept nodes and query nodes after the t-th propagation. $\mathbf{y}^c = [y_1, y_2, \cdots, y_N]^T \in \mathbb{R}^N$ and $\mathbf{y}^q = [y_1, y_2, \cdots, y_M]^T \in \mathbb{R}^M$ represent the initial score for concept nodes and query nodes. $\widetilde{\mathbf{W}}$ is the normalized $\mathbf{W}$ by $\widetilde{\mathbf{W}} = \mathbf{D}_r^{-1/2} \mathbf{W} \mathbf{D}_c^{-1/2}$, where $\mathbf{D}_r$ is a diagonal matrix with the sum of each row of $\mathbf{W}$ in the diagonal, and $\mathbf{D}_c$ is a diagonal matrix with the sum of each column of $\mathbf{W}$ in the diagonal. α is a weight ranging from 0 to 1, which represents the contribution to the updated scores by propagation.

The convergence of the sequences $\{\mathbf{f}_t^c\}$ and $\{\mathbf{f}_t^q\}$ has been proved in [5]. In another way, we can also develop a regularization framework for the above iteration algorithm. The cost function associated with $\mathbf{f}^q$ and $\mathbf{f}^c$ could be defined as

$$Q(\mathbf{f}^q, \mathbf{f}^c) =$$
$$\alpha \left(\sum_{i=1}^{|Q|} \sum_{j=1}^{|C|} W_{ij} \left(\frac{f_i^q}{\sqrt{D_Q^i}} - \frac{f_j^c}{\sqrt{D_C^j}} \right)^2 \right) + (1-\alpha)\left(\left\| \mathbf{f}^q - \mathbf{y}^q \right\|^2 + \left\| \mathbf{f}^c - \mathbf{y}^c \right\|^2 \right).$$

The above iteration algorithm could be simply obtained by differentiating $Q(\mathbf{f}^q, \mathbf{f}^c)$ with respect to $\mathbf{f}^q$ and $\mathbf{f}^c$, respectively. The first term of the right-hand side in the cost function shows that the larger w_{ij} leads to the closer of f_i^q and f_j^c. It ensures that the concepts which have stronger connection with query. It is a more intuitive explanation for the potential of BGPM in balances influences from explicit and implicit concepts.

4. EXPERIMENTAL RESULTS

We conduct video search experiments using the TRECVID 2008 data sets with 48 search tasks. Inferred average precision (infAP) is used as the evaluation criterion.

For the explicit concept detectors, we adopt the CU-VIREO374 detectors on TRECVID 2008 [10]. For the implicit concept detectors, as the previous work in [8], we use LDA to discover the latent semantic structure on SIFT bag-of-word low-level feature. However, in LDA, the predicted value in each latent topic (implicit concept) indicates the frequency; it could not be directly used as the implicit concept detector score which means probability ranged form 0 to 1. Since that, for each shot, the sum is normalized to 1; then all of the values the same latent topic are normalized to 0~1. Therefore, the normalized value could be considered as scores of the implicit concept detectors.

In the BGPM, α is set to 0.5. The initial scores for concept nodes are set to 0. The initial scores for query nodes are set to 1. Furthermore, in order to reduce the influence of the high frequency concept, we randomly choose some shots from corpus as pseudo negative samples. Since for most TRECVID 2008 search tasks, the number of relevant shots is very small compared with the whole corpus, and the random choice would not affect the retrieval performance. In the following experiment, the size of pseudo negative samples is ten times of query examples and they are added to the query nodes set with an initial score -0.1.

As the experimental design, firstly, we conduct the search tasks in two groups: one is based on query-by-example (QE); the other is based on query by example and keyword (QE+QK). Furthermore, in order to evaluate the performance of the BGPM in confusion implicit and explicit concepts, in each group, we conduct a run which only uses implicit concept nodes and query nodes in BGPM (IC_ BGPM in short), a run which only uses explicit concept nodes and query nodes in BGPM (EC_ BGPM in short), a run which averages the results of _ BGPM and EC_ BGPM (Average in short), and a run which combines implicit and explicit concept nodes in BGPM (IC+EC_ BGPM in short). Finally, the average infAP of 48 search tasks on the four runs in two groups are shown in Figure 2. Note that, the IC_BGPM result is same in the two groups, as the implicit concepts could be used only in query-by-example case.

To further study the effectiveness of the proposed method, we analyze the retrieval performance based on query types. As [4] did, we roughly group the 48 search tasks into four categories: event, person+things (PT), place, and name entity (NE) Figure 3 shows the performances of the four runs (in QE+QK group) in different query types. There is no search task that belongs to name entity type in TRECVID 2008 search task.

4.1 Explicit vs. Implicit

Obviously, as shown in Figure 2, the explicit concept-based method always outperforms the implicit concept-based one (EC_BGPM vs. IC_BGPM), even in the query-by-example group. The advantage of the supervised learning could be the main reason to explain this result. However, if we further analyze the performance of EC_BGPM and IC_BGPM on different query types as shown in Figure 3, we find that IC_BGPM surpasses EC_BGPM obviously in the Place group, which shows the potential of IC_BGPM. To further verify the potential of IC_BGPM, we take topics 0257 and 0227 as case studies. Topic 0257 is “Find shots of a plant that is the main object inside the frame area” and Topic 0227 is “Find shots of a person's face filling more than half of the frame area”. Obviously, it is impossible to find a concept form CU-VIREO374 to describe “that is the main object inside the frame area” or “filling more than half of the frame area”. It is even hard to define an explicit concept from human perspective to describe these topics. In these situations, the implicit concepts discovered from the corpus show their potential. For topics 0257 and 0227, the infAPs of IC_BGPM and EC_BGPM are 0.0151 vs. 0.0071 and 0.0624 vs. 0.0372, respectively. In addition, since we used SIFT bag-of-word as the low level feature to discover implicit concepts, it is reasonable that it likely performs well in some query for finding object or place. It also explains that why IC_BGPM has low performance in Event group. According to the above observation, we could make a conclusion that combining the explicit and implicit concepts could be a promising way to bridge the semantic gap.

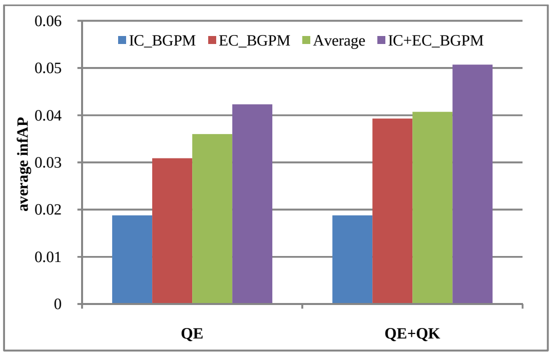

Figure 2. Performance comparison on 48 search tasks.

4.2 Average Fusion vs. BGPM

As shown in Figure 2: 1) IC+EC_BGPM is always the best run compared with IC_BGPM, EC_BGPM, and Average in different query situation. This shows the effectivity of IC+EC_BGPM in combination of implicit and explicit concepts. 2) The simply Average fusion could not adapt different query situation as IC+EC_BGPM does, where Average fusion outperforms EC_BGPM in QE group, but in QE+QK group, the improvement is very small. The following performance comparison on different query types verifies it more clearly.

As shown in Figure 3: 1) for Event group, since the performance IC_BGPM is very low, the influence of implicit should be ignored in the combination. Actually, IC+EC_BGPM successfully keep the role of explicit concepts and the final result is almost the same with EC_BGPM. Meanwhile, Average fusion fails in this situation; 2) For Person+Thing group, the performance for IG_BGPM is also much lower than EC_BGPM, but the performance of 0.0185 is not such bad, which indicates that the explicit concepts should play a key role in the combination, but note that the potential of implicit concepts should not be ignored. IC+EC_BGPM also performs well in this case which discovers the potential of implicit concepts and get an improvement from 0.0591 to 0.0704. In this case, average fusion also fails as the performance of IC_BGPM is not comparable with EC_BGPM. 3) For the Place group, IC_BGPM outperform EC_BGPM and their performances are comparable in this group, which indicates the influence from IC_BGPM and EC_BGPM should be close to equal. Since that, Average fusion finally gets an improvement in this situation. IC+EC_BGPM also does a good work, and the performance is a little better than Average fusion. Based on the performances analysis on three different query types, we could make a conclusion that BGPM could automatically balance influences of the explicit and implicit concepts according to their performances, and achieve a significantly improvement for video retrieval.

5. CONCLUSION

In this paper, firstly, we analysis two appropriate ways to bridge the semantic gap: the first one is based on the explicit concepts from human perspective (top-down) and the other one is based on the implicit concepts from computer perspective (bottom-up). We make a conclusion that combining the explicit and implicit concepts should be a promising way to bridge the semantic gap. Secondly, a Bipartite Graph Propagation Model is implemented to automatically balance influences from explicit and implicit concepts. Finally, the experimental result on TREVID 2008 video dataset verified the effectiveness of fusing the explicit and implicit concepts by BPGM to improve video retrieval.

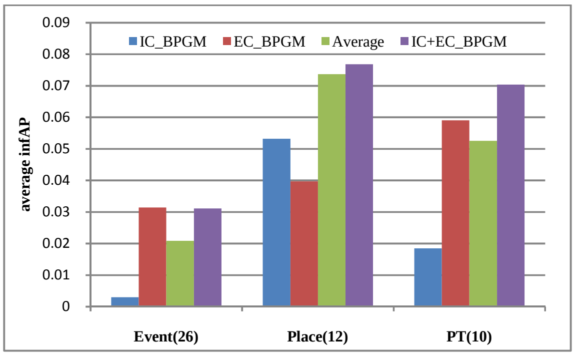

Figure 3. Performance comparison of different query types.

Currently, we only adopt SIFT bag-of-word feature to discovery the implicit concepts. In the future work, we could like to extend the low-level feature to audio and textual aspects, and try to fuse multi-modality implicit concepts by BGPM.

6. ACKNOWLEDGMENTS

This work was supported by the National Nature Science Foundation of China (60902090), National Basic Research Program of China (973Program, 2007CB311100), Co-building Program of Beijing Municipal Education Commission, Beijing New Star Project on Science & Technology (2007B071).

7. REFERENCES

[1] C. G.M. Snoek and M. Worring. Concept-Based Video Retrieval. *Foundations and Trends in Information Retrieval*: 2(4), 215-322, 2009.

[2] A. G. Hauptmann, et al. Video Retrieval Based on Semantic Concepts. *Proceedings of the IEEE* : 96(4), 602-622, 2008.

[3] A. G. Hauptmann, R. Yan, W. H. Lin, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? *IEEE Transactions on Multimedia*: 9(5), 958-966, 2007.

[4] Y.G. Jiang, C.W. Ngo, and S.F. Chang. Semantic Context Transfer across Heterogeneous Sources for Domain Adaptive

[5] X.G. Rui, M.J. Li, Z.W. Li, W.Y. Ma, Bipartite graph reinforcement model for web image annotation, In. *Proc. of ACM Multimedia*, 585-594, 2008.

[6] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[7] R. Zhao, et al. Narrowing the semantic gap-improved text-based web document retrieval using visual features. *IEEE Transactions on Multimedia:* 4(2), 189-200, 2002

[8] J. Cao, Y.D. Zhang, B.L. Feng, X.F Hua, L. Bao, X. Zhang and J.T. Li. MCG-ICT-CAS TRECVID2008 Search Task Report. In. *Proc. of TRECVID Workshop*, 2008.

[9] R. Yan, A. G. Hauptmann. Probabilistic Latent Query Analysis for Combining Multiple Retrieval Sources. In *Proc. of ACM SIGIR*, 324-331, 2006

[10] Y.G. Jiang, A. Yanagawa, S.F. Chang, and C.W. Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection, *Columbia University ADVENT Technical Report #223-2008-1*, Aug. 2008.